Mathématiques pour Informaticiens

Ernst Hairer

Université de Genève
Section de mathématiques
Case postale 240
CH-1211 Genève 24

Juin 2004

Contents

Remerciements

Ce polycopié accompagne le cours “Mathématiques pour informaticiens” (4 heures par semaine et 2 heures d'exercices) donné au semestre d'été 2004.

Chapter I

Topologie de $\mathbb{R}^n$ et fonctions continues

L'étude des fonctions d'*une* variable réelle est un des sujets du cours "Analyse I" (semestre d'hiver). Dans ce chapitre, nous discutons des notions qui permetteront une généralisation aux fonctions de *plusieurs* variables. De telles fonctions apparaissent partout en pratique. Par exemple, la température dans une salle est une fonction qui dépend de l'espace (trois coordonnées) et du temps. Elle est donc une fonction de quatre variables.

Ce chapitre suit de près la présentation des paragraphes IV.1 et IV.2 du livre "L'analyse au fil de l'histoire" de Hairer & Wanner (Springer-Verlag 2002). Pour plus d'informations (remarques historiques, démonstrations détaillées, ...), la lecture de ce livre est vivement conseillée.

I.1 Distances et normes

Nous considérons des couples (x_1, x_2) de nombres réels, et des n-uples $(x_1, x_2, \ldots, x_n)$. L'ensemble de tous les couples est

$$\mathbb{R}^2 = \mathbb{R} \times \mathbb{R} = \{(x_1, x_2) \; ; \; x_1, x_2 \in \mathbb{R}\} \tag{1.1}$$

et l'ensemble de tous les n-uples est

$$\mathbb{R}^n = \mathbb{R} \times \mathbb{R} \times \ldots \times \mathbb{R} = \{(x_1, x_2, \ldots, x_n) \; ; \; x_k \in \mathbb{R}, \; k = 1, \ldots, n\}. \tag{1.2}$$

Avec l'addition (composante par composante) et avec la multiplication par des nombres réels, cet ensemble devient un espace vectoriel (voir le chapitre II du cours "Algèbre I", semestre d'hiver). Les éléments de $\mathbb{R}^n$ sont donc des vecteurs. Dans ce chapitre, nous ne distinguons pas les vecteurs colonnes et les vecteurs lignes.

Géométriquement, l'espace $\mathbb{R}^2$ peut être interprété comme un plan; les composantes x_1 et x_2 étant les coordonnées cartésiennes. Par le théorème de Pythagore, la distance $d(x, y)$ entre deux points $x = (x_1, x_2)$ et $y = (y_1, y_2)$ est donnée par (figure I.1, gauche)

$$d(x, y) = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2}. \tag{1.3}$$

On voit que cette distance ne dépend que de la différence $y - x$. Ceci justifie l'écriture $\|y - x\|_2$, où $\|z\|_2 = \sqrt{z_1^2 + z_2^2}$ si $z = (z_1, z_2)$.

Pour calculer la distance entre $x = (x_1, x_2, x_3)$ et $y = (y_1, y_2, y_3)$ dans l'espace $\mathbb{R}^3$, nous appliquons deux fois le théorème de Pythagore (d'abord au triangle DEF et ensuite à ABC, voir figure I.1, droite) et nous obtenons $d(x, y) = \|y - x\|_2$, où $\|z\|_2 = \sqrt{z_1^2 + z_2^2 + z_3^2}$.

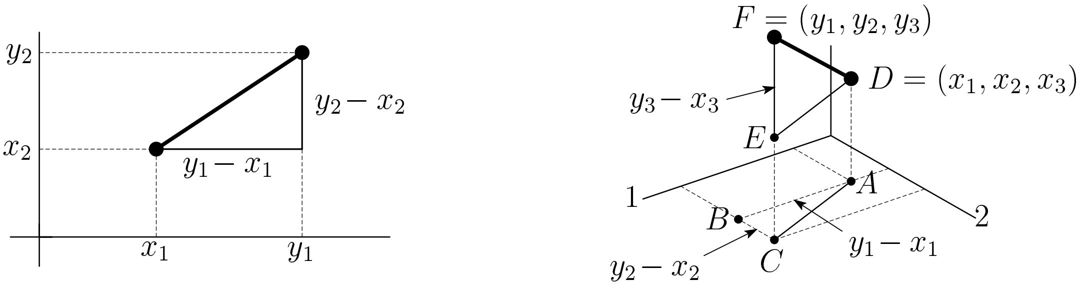


Figure I.1: Distances dans $\mathbb{R}^2$ et dans $\mathbb{R}^3$

Dans l'espace $\mathbb{R}^n$ de dimension n, nous définissons par analogie

$$\|z\|_2 = \sqrt{z_1^2 + z_2^2 + \ldots + z_n^2}, \tag{1.4}$$

et nous appelons cette expression la *norme euclidienne* de $z = (z_1, z_2, \ldots, z_n)$. La distance entre $x \in \mathbb{R}^n$ et $y \in \mathbb{R}^n$ est alors donnée par $d(x, y) = \|y - x\|_2$.

Théorème 1.1 *La norme euclidienne (1.4) satisfait :*

(N1) $\|x\| \geq 0$ *et* $\|x\| = 0 \Leftrightarrow x = 0$,

(N2) $\|\lambda x\| = |\lambda| \cdot \|x\|$ *pour* $\lambda \in \mathbb{R}$,

(N3) $\|x + y\| \leq \|x\| + \|y\|$ *(inégalité du triangle).*

Démonstration. Les propriétés (N1) et (N2) sont triviales. La démonstration de (N3) est basée sur l'inégalité de Cauchy-Schwarz

$$|\langle x, y\rangle| \leq \|x\| \cdot \|y\|, \tag{1.5}$$

où $\langle x, y\rangle := \sum_{k=1}^n x_k y_k$ désigne le produit scalaire de deux vecteurs x et y (voir le chapitre VI du cours "Algèbre I"). L'inégalité du triangle est maintenant une conséquence de

$$\|x+y\|^2 = \langle x+y, x+y\rangle = \|x\|^2 + 2\langle x, y\rangle + \|y\|^2 \leq \|x\|^2 + 2\|x\| \cdot \|y\| + \|y\|^2 = (\|x\| + \|y\|)^2.$$

□

Par la suite, nous utiliserons très rarement la formule explicite de l'équation (1.4). Souvent, il est confortable d'utiliser d'autres expressions satisfaisant les propriétés (N1), (N2) et (N3).

Définition 1.2 (norme) *Une norme sur $\mathbb{R}^n$ est une application $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ satisfaisant (N1), (N2) et (N3). L'espace $\mathbb{R}^n$ muni d'une norme s'appelle un espace normé.*

Exemples. En plus de la norme euclidienne (1.4), nous considérons

$$\|x\|_1 = \sum_{k=1}^n |x_k| \qquad \text{norme } \ell_1, \tag{1.6}$$

$$\|x\|_\infty = \max_{k=1,\ldots,n} |x_k| \qquad \text{norme maximum.} \tag{1.7}$$

La vérification des propriétés (N1), (N2) et (N3) pour ces normes est facile. La notation (c.-à-d. les indices $1, 2, \infty$) n'est pas choisie par hasard. En effet, les trois normes sont des cas particuliers de

$$\|x\|_p = \Big(\sum_{k=1}^n |x_k|^p\Big)^{1/p} \qquad \text{norme } \ell_p, \quad 1 \leq p < \infty \tag{1.8}$$

et $\|x\|_\infty = \lim_{p\to\infty} \|x\|_p$.

Théorème 1.3 *Pour tout $x \in \mathbb{R}^n$, on a*

$$\|x\|_\infty \le \|x\|_2 \le \|x\|_1 \le n \cdot \|x\|_\infty. \tag{1.9}$$

Démonstration. Nous ne démontrons que la deuxième inégalité. En prenant le carré $\|x\|_1^2$ dans l'équation (1.6), nous obtenons la somme des carrés $\sum x_k^2$ (c'est-à-dire $\|x\|_2^2$) et les produits mixtes $|x_k| \cdot |x_l|$, tous non négatifs. Ceci implique que $\|x\|_1^2 \ge \|x\|_2^2$. □

Ce résultat montre que les normes $\|x\|_1$, $\|x\|_2$ et $\|x\|_\infty$ sont équivalentes dans l'esprit de la définition suivante.

Définition 1.4 (équivalence de normes) *On dit que deux normes $\|\cdot\|_p$ et $\|\cdot\|_q$ sont équivalentes s'il existe des constantes positives C_1 et C_2 telles que*

$$C_1 \|x\|_p \le \|x\|_q \le C_2 \|x\|_p \qquad \textit{pour tout} \quad x \in \mathbb{R}^n. \tag{1.10}$$

I.2 Convergence de suites de vecteurs

Au cours "Analyse I", nous avons vu des suites de nombres réels ainsi que les notions de convergence, de suites de Cauchy, etc. Il s'agit maintenant d'étendre ces définitions et résultats à des suites de vecteurs. Nous considérons donc $\{x_i\}_{i\ge 1}$, où chaque x_i est un vecteur, c.-à-d.

$$x_i = (x_{1i}, x_{2i}, \ldots, x_{ni}), \qquad i = 1, 2, 3, \ldots\ . \tag{2.1}$$

Définition 2.1 (convergence de suites de vecteurs) *On dit que la suite $\{x_i\}_{i\ge 1}$, donnée par (2.1), converge vers le vecteur $a = (a_1, a_2, \ldots, a_n) \in \mathbb{R}^n$ si*

$$\forall \varepsilon > 0 \quad \exists N \ge 1 \quad \forall i \ge N \qquad \|x_i - a\| < \varepsilon.$$

Comme pour des suites dans $\mathbb{R}$, nous écrivons $x_i \to a$ ou bien $\lim_{i\to\infty} x_i = a$.

La seule différence par rapport à la définition pour des suites de nombres réels est que les "valeurs absolues" sont remplacées par des "normes".

La figure I.2 montre la suite $x_i = (0.6 + 0.56\, r^i \cos(0.6 + 0.7\, i)\,,\, 0.4 + 0.4\, r^i \sin(0.6 + 0.7\, i))$ dans $\mathbb{R}^2$ avec $r = 0.86$. Elle converge vers $a = (0.6, 0.4)$.

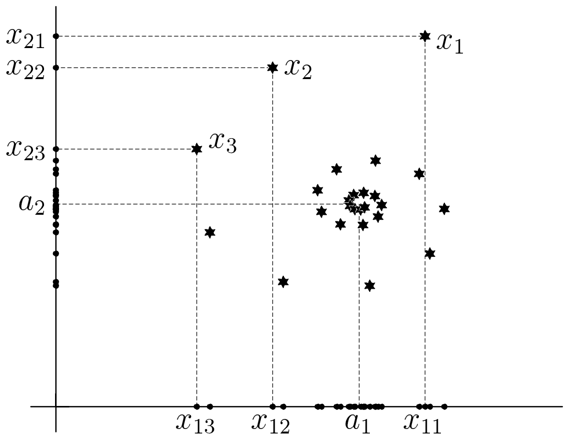


Figure I.2: Illustration d'une suite convergente dans $\mathbb{R}^2$

Attention Dans la définition 2.1, nous n'avons pas encore précisé quelle norme il faut prendre. Nous observons que si $\|\cdot\|_p$ est équivalente à $\|\cdot\|_q$, alors on a

$$\text{convergence avec } \|\cdot\|_p \quad \Longleftrightarrow \quad \text{convergence avec } \|\cdot\|_q. \tag{2.2}$$

En effet, $\|x_i - a\|_p < \varepsilon$ et (1.10) impliquent que $\|x_i - a\|_q < C_2\varepsilon$. Puisque $\varepsilon > 0$ est arbitraire dans la définition 2.1, nous pouvons le remplacer par $\varepsilon' = C_2\varepsilon$, et nous voyons que la convergence avec $\|\cdot\|_p$ implique celle avec $\|\cdot\|_q$.

Nous savons déjà (théorème 1.3) que $\|\cdot\|_1$, $\|\cdot\|_2$ et $\|\cdot\|_\infty$ sont équivalentes ; nous verrons plus loin (théorème 4.6) que toutes les normes dans $\mathbb{R}^n$ sont équivalentes. Par conséquent, on peut se servir de n'importe quelle norme dans la définition 2.1.

Théorème 2.2 (critère de convergence) *Pour une suite de vecteurs (2.1), on a*

$$\lim_{i\to\infty} x_i = a \qquad \Longleftrightarrow \qquad \lim_{i\to\infty} x_{ki} = a_k \quad \textit{pour} \quad k = 1, 2, \ldots, n,$$

i.e. la convergence dans $\mathbb{R}^n$ est équivalente à la convergence composante par composante.

Démonstration. Prenons la norme maximum (1.7) pour laquelle

$$\|x_i - a\|_\infty < \varepsilon \quad \Longleftrightarrow \quad |x_{ki} - a_k| < \varepsilon \;\text{ pour }\; k = 1, 2, \ldots, n.$$

Avec cette norme dans la définition 2.1, le résultat est immédiat. □

Avec l'observation qu'il faut seulement remplacer "valeurs absolues" par "normes", on peut étendre beaucoup de définitions et résultats du cours "Analyse I" à une dimension supérieure. Par exemple, nous disons qu'une suite des vecteurs $\{x_i\}_{i\geq 1}$ est *bornée,* s'il existe un nombre $B \geq 0$ tel que $\|x_i\| \leq B$ pour tout $i \geq 1$. De nouveau, la propriété d'être bornée ne dépend pas de la norme choisie. Comme dans $\mathbb{R}$, nous voyons que toute suite convergente est bornée.

On dit qu'une suite $\{x_i\}_{i\geq 1}$ est une *suite de Cauchy* si

$$\forall \varepsilon > 0 \quad \exists N \geq 1 \quad \forall i \geq N \quad \forall \ell \geq 1 \quad \|x_i - x_{i+\ell}\| < \varepsilon. \tag{2.3}$$

En utilisant la norme maximum dans (2.3), on constate que cette définition est équivalente au fait que, pour $k = 1, \ldots, n$, les suites réelles $\{x_{ki}\}_{i\geq 1}$ sont des suites de Cauchy. Par conséquent, nous obtenons directement la généralisation du critère de Cauchy.

Théorème 2.3 (critère de Cauchy dans $\mathbb{R}^n$) *Une suite de vecteurs dans $\mathbb{R}^n$ est convergente si et seulement si elle est une suite de Cauchy.* □

Le théorème de Bolzano-Weierstrass reste aussi vrai dans $\mathbb{R}^n$. Sa démonstration, par contre, nécessite quelques idées supplémentaires.

Théorème 2.4 (Bolzano-Weierstrass dans $\mathbb{R}^n$) *Chaque suite bornée de vecteurs dans $\mathbb{R}^n$ admet une sous-suite convergente.*

Démonstration. Soit $\{x_i\}_{i\geq 1}$ une suite bornée dans $\mathbb{R}^n$. La suite de ses premières composantes $\{x_{1i}\}_{i\geq 1}$ est bornée et on peut appliquer le théorème de Bolzano-Weierstrass dans $\mathbb{R}$. Elle admet alors une sous-suite convergente, disons

$$x_{1,1},\; x_{1,5},\; x_{1,9},\; x_{1,22},\; x_{1,37},\; x_{1,58},\; x_{1,238},\; x_{1,576},\; \ldots\; . \tag{2.4}$$

Passons aux deuxièmes composantes. L'idée essentielle est de ne considérer que celles qui correspondent à la sous-suite (2.4) et non pas celles de la suite toute entière. Cette suite est bornée, et nous pouvons de nouveau appliquer le théorème de Bolzano-Weierstrass dans $\mathbb{R}$ pour obtenir une sous-suite convergente, disons,

$$x_{2,1},\ x_{2,9},\ x_{2,58},\ x_{2,576},\ \ldots\ . \tag{2.5}$$

Maintenant, les premières et les deuxièmes composantes de la suite $x_1, x_9, x_{58}, x_{576}, \ldots$ convergent. Pour $n = 2$, la démonstration est terminée. Pour $n > 2$, nous examinons les troisièmes composantes avec des indices correspondant à (2.5), et ainsi de suite. Après n étapes, il reste une suite dont toutes les composantes convergent. □

I.3 Voisinages, ensembles ouverts et fermés

Pour des ensembles A et B dans $\mathbb{R}^n$, nous utilisons les notations

$$\begin{aligned}
A \subset B &\quad \text{si les éléments de } A \text{ appartiennent à } B \text{ (sous-ensemble)},\\
A \cap B &= \{x \in \mathbb{R}^n\ ;\ x \in A \quad \text{et} \quad x \in B\} \quad \text{(intersection)},\\
A \cup B &= \{x \in \mathbb{R}^n\ ;\ x \in A \quad \text{ou} \quad x \in B\} \quad \text{(réunion ou union)},\\
A \setminus B &= \{x \in \mathbb{R}^n\ ;\ x \in A \quad \text{mais} \quad x \notin B\} \quad \text{(différence)},\\
\complement A &= \{x \in \mathbb{R}^n\ ;\ x \notin A\} \quad \text{(complémentaire)}.
\end{aligned}$$

Le rôle de l'intervalle ouvert est joué par

$$B_\varepsilon(a) = \{x \in \mathbb{R}^n\ ;\ \|x - a\| < \varepsilon\}, \tag{3.1}$$

appelé *disque* (ou *boule*) de rayon ε et de centre a (voir la figure I.3).

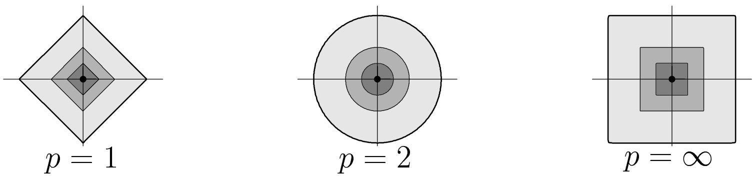
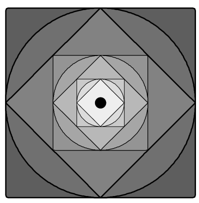


Figure I.3: Disques de rayon $\varepsilon = 1,\ 1/2,\ 1/4$ pour $\|x\|_1$, $\|x\|_2$ et $\|x\|_\infty$,

Définition 3.1 (voisinage) *Soit $a \in \mathbb{R}^n$ donné. Un voisinage de a est un ensemble $V \subset \mathbb{R}^n$ qui contient un disque de rayon positif centré en a, i.e.,*

$$V \text{ est voisinage de } a \qquad \Longleftrightarrow \qquad \exists \varepsilon > 0 \quad B_\varepsilon(a) \subset V.$$

Le disque $B_\varepsilon(a)$ dépend de la norme utilisée ($\|\cdot\|_1, \|\cdot\|_2$ ou $\|\cdot\|_\infty, \ldots$) ; la définition d'un voisinage, par contre, est *indépendante* de la norme utilisée, à condition que les normes soient équivalentes. Chaque $B_\varepsilon(a)$ pour une norme contient un $B_{\varepsilon'}(a)$ pour l'autre norme (voir le dessin à coté).



Définition 3.2 (ensemble ouvert) *Un ensemble $U \subset \mathbb{R}^n$ est ouvert s'il est un voisinage de chacun de ses points, i.e.*

$$U \text{ ouvert} \qquad \Longleftrightarrow \qquad \forall x \in U \quad \exists \varepsilon > 0 \quad B_\varepsilon(x) \subset U.$$

Définition 3.3 (ensemble fermé) *Un ensemble $F \subset \mathbb{R}^n$ est fermé si chaque suite convergente $\{x_i\}_{i\geq 1}$ avec $x_i \in F$ a sa limite dans F, i.e.*

$$F \text{ fermé} \quad \Longleftrightarrow \quad a = \lim_{i\to\infty} x_i \quad \text{et} \quad x_i \in F \quad \text{impliquent} \quad a \in F.$$

Exemples dans $\mathbb{R}$.

L'intervalle dit "ouvert" $(a,b) = \{x \in \mathbb{R}\ ;\ a < x < b\}$ est un ensemble ouvert. En effet, pour un $x \in (a,b)$, le nombre $\varepsilon := \min(x-a, b-x)$ est strictement positif, et nous avons $B_\varepsilon(x) \subset (a,b)$. Par contre, la suite $\{a + 1/i\}$ (pour $i \geq 1$) est convergente, ses éléments sont dans (a,b) à partir d'un certain i, mais la limite n'est pas dans (a,b). Donc, l'ensemble (a,b) n'est pas fermé.

L'ensemble dit "fermé" $[a,b] = \{x \in \mathbb{R}\ ;\ a \leq x \leq b\}$ est fermé (faire la démonstration à l'aide d'un théorème du cours "Analyse I"). Cependant, ni a ni b n'ont un voisinage entièrement inclus dans $[a,b]$. L'intervalle $[a,b]$ n'est donc pas ouvert.

L'intervalle $A = [a,b)$ n'est ni ouvert ni fermé, puisque a n'a pas de voisinage dans $[a,b)$ et la limite de la suite convergente $\{b - 1/i\}$ n'est pas dans $[a,b)$.

Enfin, l'ensemble $\mathbb{R} = (-\infty, +\infty)$ est ouvert et fermé, ainsi que l'ensemble vide $\emptyset$.

Lemme 3.4 *Soit $\|\cdot\|$ une norme arbitraire de $\mathbb{R}^n$.*

a) *L'ensemble* $A = \{x \in \mathbb{R}^n\ ;\ \|x\| < 1\}$ *est ouvert.*

b) *L'ensemble* $A = \{x \in \mathbb{R}^n\ ;\ \|x\| \leq 1\}$ *est fermé.*

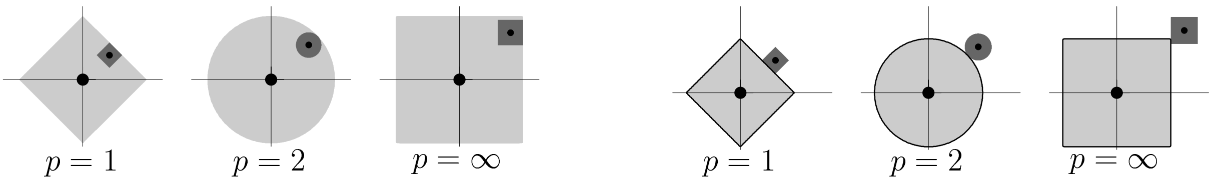


Figure I.4: Ensembles ouverts $\{x\,;\,\|x\|_p < 1\}$ (gauche) et fermés $\{x\,;\,\|x\|_p \leq 1\}$ (droite)

Démonstration. a) Pour $a \in A$ prenons $\varepsilon = 1 - \|a\|$ qui est positif. Avec ce choix, nous avons $B_\varepsilon(a) \subset A$ (voir figure I.4, gauche). En effet, par l'inégalité du triangle, nous avons pour $x \in B_\varepsilon(a)$

$$\|x\| = \|x - a + a\| \leq \|x - a\| + \|a\| < \varepsilon + \|a\| = 1.$$

Donc, A est ouvert.

b) Considérons une suite $\{x_i\}_{i\geq 1}$ vérifiant $x_i \in A$ (pour tout i) et convergeant vers a. L'inégalité de triangle implique

$$\|a\| = \|x_i - x_i + a\| \leq \|x_i\| + \|x_i - a\| \leq 1 + \|x_i - a\| < 1 + \varepsilon$$

pour $i \geq N$. Ceci est vrai pour tout $\varepsilon > 0$. Par conséquent, $\|a\| \leq 1$ et A est fermé. □

Autres exemples.

Le demi-plan $A = \{x \in \mathbb{R}^2\ ;\ x_1 + x_2 > 2\}$ est un ensemble ouvert; par contre, le demi-plan $A = \{x \in \mathbb{R}^2\ ;\ x_1 + x_2 \geq 2\}$ est un fermé (même démonstrations que tout à l'heure). On verra plus tard que l'ensemble $A = \{x \in \mathbb{R}^2\ ;\ f(x_1,x_2) > 0\}$ est ouvert et que $A = \{x \in \mathbb{R}^2\ ;\ f(x_1,x_2) \geq 0\}$ est fermé si $f(x_1,x_2)$ est une fonction continue.

L'ensemble $A = \{x \in \mathbb{R}^2\ ;\ x_1,\, x_2 \in \mathbb{Q},\ \|x\| \leq 1\}$ n'est ni ouvert ni fermé, car chaque disque contient des points irrationnels; et une limite de points rationnels peut être irrationnelle.

Le célèbre ensemble de Cantor (figure I.5) est donné par

$$A = [0,1] \setminus \{(1/3,2/3) \cup (1/9,2/9) \cup (7/9,8/9) \cup \ldots\} = \Big\{ x \; ; \; x = \sum_{i=1}^{\infty} a_i 3^{-i} \; , \; a_i \in \{0,2\} \Big\}. \tag{3.2}$$

Cet ensemble n'est pas ouvert (par ex. $x = 1/3$ n'a pas de voisinage dans A) mais il est fermé (voir la remarque après la démonstration du théorème 3.6).



Figure I.5: Ensemble de Cantor

Le *triangle de Sierpiński* et le *tapis de Sierpiński* (figure I.6) sont des généralisations bidimensionnelles de l'ensemble de Cantor. Ces dessins ne nous plaisent pas seulement par leur beauté, mais nous rappellent que les ensembles peuvent être des objets bien compliqués.

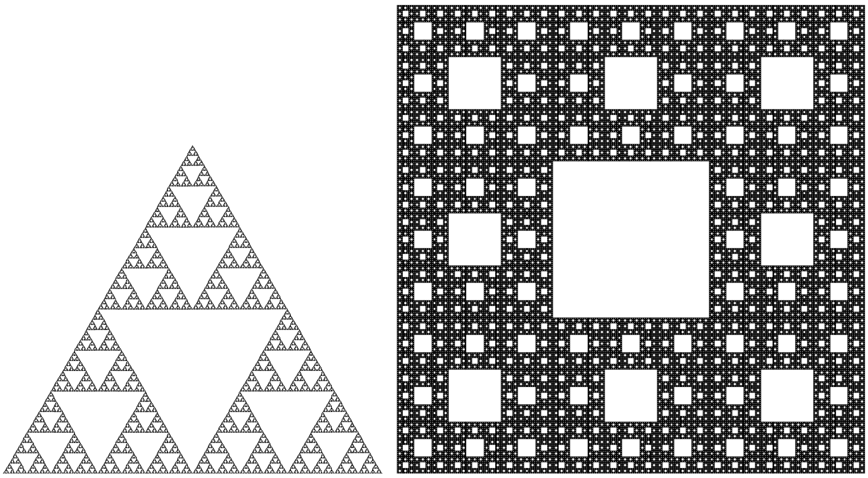


Figure I.6: Triangle de Sierpiński et tapis de Sierpiński

Théorème 3.5 *On a*

$$\text{i)} \quad F \text{ fermé} \implies \complement F \text{ ouvert,}$$
$$\text{ii)} \quad U \text{ ouvert} \implies \complement U \text{ fermé.}$$

Démonstration. i) Supposons que $\complement F$ ne soit pas ouvert. Il existe alors un $a \in \complement F$ (i.e. $a \notin F$) tel que, pour tout $\varepsilon > 0$, on ait $B_\varepsilon(a) \not\subset \complement F$. En prenant $\varepsilon = 1/i$, nous pouvons choisir une suite $\{x_i\}_{i\geq 1}$ satisfaisant $x_i \in F$ et $\|x_i - a\| < 1/i$. Comme F est fermé, nous obtenons $a \in F$, d'où une contradiction.

ii) Supposons que $\complement U$ ne soit pas fermé. Il existe alors une suite $x_i \in \complement U$ (i.e. $x_i \notin U$) convergeant vers un $a \notin \complement U$, (i.e. $a \in U$). Comme U est ouvert, nous avons $B_\varepsilon(a) \subset U$ pour un certain $\varepsilon > 0$. Par conséquent, $x_i \notin B_\varepsilon(a)$ pour tout i, d'où une contradiction. □

Théorème 3.6 *Pour un nombre fini d'ensembles, on a*

i) $U_1, U_2, \ldots, U_m$ *ouverts* $\implies$ $U_1 \cap U_2 \cap \ldots \cap U_m$ *est ouvert,*

ii) $F_1, F_2, \ldots, F_m$ *fermés* $\implies$ $F_1 \cup F_2 \cup \ldots \cup F_m$ *est fermé.*

Pour une famille arbitraire d'ensembles (indexée par l'ensemble Λ), on a

iii) U_λ *ouvert pour tout* λ $\Rightarrow$ $\bigcup_{\lambda\in\Lambda} U_\lambda = \{x \in \mathbb{R}^n ; \exists \lambda \in \Lambda, \; x \in U_\lambda\}$ *est ouvert,*

iv) F_λ *fermé pour tout* λ $\Rightarrow$ $\bigcap_{\lambda\in\Lambda} F_\lambda = \{x \in \mathbb{R}^n ; \forall \lambda \in \Lambda, \; x \in F_\lambda\}$ *est fermé.*

Démonstration. Commençons par la démonstration de (i). Soit $x \in U_1 \cap \ldots \cap U_m$, donc $x \in U_k$ pour tout $k = 1, \ldots, m$. Comme U_k est ouvert, il existe un $\varepsilon_k > 0$ tel que $B_{\varepsilon_k}(x) \subset U_k$. Prenons $\varepsilon = \min(\varepsilon_1, \ldots, \varepsilon_m)$; alors $\varepsilon > 0$ et $B_\varepsilon(x) \subset U_1 \cap \ldots \cap U_m$.

La démonstration de (iii) est encore plus simple ; nous l'omettons. Les équivalences (i) $\Leftrightarrow$ (ii) et (iii) $\Leftrightarrow$ (iv) sont une conséquence des "règles de Morgan"

$$\complement(U_1 \cap U_2) = (\complement U_1) \cup (\complement U_2), \qquad \complement(U_1 \cup U_2) = (\complement U_1) \cap (\complement U_2),$$

et du théorème 3.5. □

Remarque. Ce théorème nous permet de démontrer que l'ensemble de Cantor défini par (3.2) est fermé. En effet, son complémentaire

$$\complement A = (-\infty, 0) \cup (1, \infty) \cup (1/3, 2/3) \cup (1/9, 2/9) \cup (7/9, 8/9) \cup \ldots$$

est une union infinie d'intervalles ouverts. Le théorème 3.6 implique donc que A est fermé.

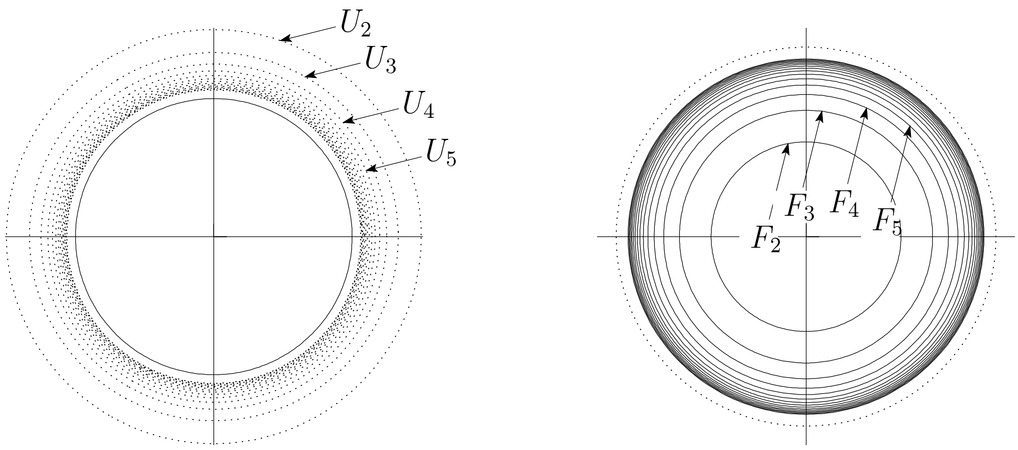


Figure I.7: Ensembles ouverts d'intersection fermée (gauche) et ensembles fermés d'union ouverte (droite)

Remarque. Les affirmations (i) et (ii) du théorème 3.6 ne sont, en général, pas vraies pour un nombre infini d'ensembles.

Considérons par exemple la famille d'ensembles ouverts

$$U_i = \Big\{x \in \mathbb{R}^2 \; ; \; \|x\| < 1 + 1/i\Big\}, \tag{1.26}$$

dont l'intersection $U_2 \cap U_3 \cap U_4 \cap \ldots = \{x \in \mathbb{R}^2 \; ; \; \|x\| \leq 1\}$ n'est pas ouverte (figure I.7, gauche).

De façon analogue, la famille d'ensembles fermés (figure I.7, droite)

$$F_i = \Big\{x \in \mathbb{R}^2 \; ; \; \|x\| \leq 1 - 1/i\Big\}, \tag{1.27}$$

a une union $F_2 \cup F_3 \cup F_4 \cup \ldots = \{x \in \mathbb{R}^2 \; ; \; \|x\| < 1\}$ qui n'est pas fermée.

I.4 Fonctions continues

Soit A un sous-ensemble de $\mathbb{R}^n$. Une fonction

$$f : A \to \mathbb{R}^m \tag{4.1}$$

envoie le vecteur $x = (x_1, \ldots, x_n) \in A$ sur le vecteur $y = (y_1, \ldots, y_m) \in \mathbb{R}^m$. Chaque composante de y est une fonction de n variables. Nous écrivons alors

$$y = f(x) \qquad \text{ou} \qquad \begin{array}{rcl} y_1 & = & f_1(x_1, \ldots, x_n) \\ & \vdots & \\ y_m & = & f_m(x_1, \ldots, x_n). \end{array} \tag{4.2}$$

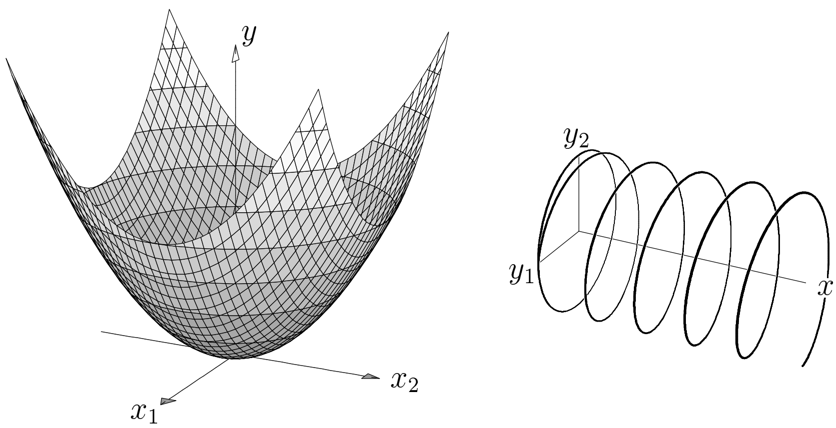


Figure I.8: Paraboloïde (gauche) et hélice (droite)

Exemples

a) Une fonction $(m = 1)$ de deux variables $(n = 2)$ peut être interprétée comme une surface dans $\mathbb{R}^3$. Par exemple, la fonction $y = x_1^2 + x_2^2$ représente un paraboloïde (figure I.8, gauche).

b) Deux fonctions $(m = 2)$ d'une variable $(n = 1)$ représentent une courbe dans $\mathbb{R}^3$. Par exemple, l'hélice de la figure I.8 (droite) est donnée par $y_1 = \cos 10x$, $y_2 = \sin 10x$. Si nous projetons la courbe sur le plan (y_1, y_2), nous obtenons une "représentation paramétrique" d'une courbe dans $\mathbb{R}^2$ (un cercle dans cet exemple).

Définition 4.1 (continuité) *Une fonction $f : A \to \mathbb{R}^m$, $A \subset \mathbb{R}^n$ est continue en $x_0 \in A$ si*

$$\forall \varepsilon > 0 \quad \exists \delta > 0 \quad \forall x \in A \; : \; \|x - x_0\| < \delta \qquad \|f(x) - f(x_0)\| < \varepsilon.$$

C'est précisément la définition de la continuité d'une fonction à une variable avec les valeurs absolues remplacées par des normes. Cette définition ne dépend pas des normes choisies, pourvu qu'elles soient équivalentes. En utilisant la norme maximum dans $\mathbb{R}^m$, nous obtenons le résultat suivant (à comparer avec le théorème 2.2).

Théorème 4.2 *Une fonction $f : A \to \mathbb{R}^m$, $A \subset \mathbb{R}^n$ donnée par (4.2) est continue en $x_0 \in A$ si et seulement si chaque fonction $f_j : A \to \mathbb{R}$ est continue en x_0 ($j = 1, \ldots, m$).* □

Une conséquence de ce théorème est qu'il suffit de considérer le cas $m = 1$ pour l'étude de la continuité.

Il est évident qu'une fonction constante $f(x) = c$ est partout continue. La projection de $x = (x_1, \ldots, x_n)$ sur sa $k$ième coordonnée, i.e. $p(x) = x_k$, est également continue en chaque point $x_0 = (x_{10}, \ldots, x_{n0})$, puisque $|x_k - x_{k0}| \leq \|x - x_0\|$ (choisir $\delta = \varepsilon$ dans la définition 4.1).

Comme pour des fonctions à une variable, on démontre que le produit et le quotient (si le dénominateur est non nul) de fonctions continues est continue. Par conséquent, les polynômes en plusieurs variables, par exemple $f(x_1, x_2, x_3) = x_1^4 x_2^5 - x_1 x_2^3 x_3 + 4x_2^5 - 1$, sont partout continus; et les fonctions rationnelles sont continues aux points où le dénominateur est non nul.

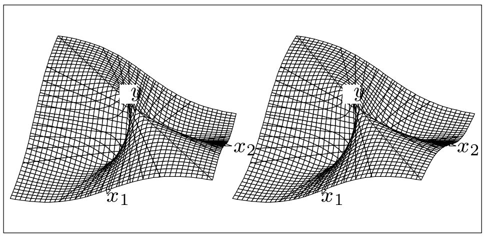


Figure I.9: Stéréogramme de la fonction discontinue $f(x_1, x_2)$ de la formule (4.3) (tenir le dessin à 20 cm des yeux et "loucher" à travers le papier vers un objet situé 20 cm derrière. Alors, les deux dessins fusionnent et un dessin 3D apparaît.)

Exemple 4.3 Considérons la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ donnée par

$$f(x_1, x_2) = \begin{cases} \dfrac{x_1 x_2}{x_1^2 + x_2^2} & \text{si} \quad x_1^2 + x_2^2 > 0 \\ 0 & \text{si} \quad x_1 = x_2 = 0 \end{cases} \tag{4.3}$$

(voir figure I.9). Comme elle est une fonction rationnelle, elle est certainement continue aux points vérifiant $x_1^2 + x_2^2 > 0$. Pour expliquer son comportement près de l'origine, utilisons les coordonnées polaires $x_1 = r\cos\varphi$, $x_2 = r\sin\varphi$; il vient (pour $r > 0$)

$$f(r\cos\varphi\,,\ r\sin\varphi) = \frac{r^2\cos\varphi\sin\varphi}{r^2} = \frac{1}{2}\sin 2\varphi.$$

Par conséquent, la fonction est constante sur les droites passant par l'origine, et cette constante dépend de l'angle φ. Dans tout voisinage de $(0, 0)$, la fonction (4.3) prend toutes les valeurs entre $+1/2$ et $-1/2$. Elle ne peut donc pas être continue en $(0, 0)$.

Définition 4.4 (ensemble compact) *Pour un ensemble $K \subset \mathbb{R}^n$, on dit que*

$$K \text{ est compact} \qquad \Longleftrightarrow \qquad K \text{ est borné et fermé.}$$

Beaucoup de résultats pour des fonctions continues à une variable possèdent une généralisation à plusieurs variables. Un résultat qu'on obtient de nouveau en remplaçant "valeurs absolues" par "normes" est le suivant.

Théorème 4.5 *Soit $K \subset \mathbb{R}^n$ un ensemble compact et soit $f : K \to \mathbb{R}$ continue sur K. Alors, f est bornée sur K et admet un maximum et un minimum, i.e. il existe $u \in K$ et $U \in K$ tels que*

$$f(u) \leq f(x) \leq f(U) \qquad \text{pour tout} \quad x \in K. \qquad \square$$

Ce théorème conduit au résultat suivant, déjà annoncé au début de ce chapitre.

Théorème 4.6 (équivalence de normes) *Toutes les normes dans $\mathbb{R}^n$ sont équivalentes. C'est-à-dire que si $N : \mathbb{R}^n \to \mathbb{R}$ est une application satisfaisant les conditions (N1), (N2) et (N3) du théorème 1.1, i.e.*

(N1) $\quad N(x) \geq 0 \quad$ *et* $\quad N(x) = 0 \Leftrightarrow x = 0,$

(N2) $\quad N(\lambda x) = |\lambda|\, N(x) \quad$ *pour* $\quad \lambda \in \mathbb{R},$

(N3) $\quad N(x+y) \leq N(x) + N(y) \quad$ *(inégalité du triangle),*

alors il existe des constantes $C_1 > 0$ et $C_2 > 0$ telles que

$$C_1\, \|x\|_2 \leq N(x) \leq C_2\, \|x\|_2 \qquad \textit{pour tout} \quad x \in \mathbb{R}^n. \tag{4.4}$$

Démonstration. a) Écrivons $x = x_1e_1 + x_2e_2 + \ldots + x_ne_n$, dans la base canonique $e_1 = (1,0,\ldots,0)$, $e_2 = (0,1,0,\ldots,0)$, etc. On déduit de (N3), (N2) et de l'inégalité de Cauchy-Schwarz (1.5) que

$$\begin{aligned} N(x) &= N(x_1e_1 + \ldots + x_ne_n) \leq N(x_1e_1) + \ldots + N(x_ne_n) \\ &\leq |x_1| \cdot N(e_1) + \ldots + |x_n| \cdot N(e_n) \leq \|x\|_2 \cdot C_2, \end{aligned} \tag{4.5}$$

avec $C_2 = \sqrt{N(e_1)^2 + \ldots + N(e_n)^2}$. La deuxième inégalité de (4.4) est ainsi démontrée.

b) Montrons ensuite que $N(x)$ est continue (avec la norme euclidienne dans la définition 4.1). On a

$$N(x) - N(x_0) = N(x - x_0 + x_0) - N(x_0) \leq N(x - x_0) + N(x_0) - N(x_0) \leq C_2\|x - x_0\|_2$$

et, de la même façon, $N(x_0) - N(x) = \ \ldots \ \leq C_2\|x_0 - x\|_2$. On en déduit que

$$|N(x) - N(x_0)| \leq C_2\|x - x_0\|_2$$

ce qui implique la continuité de $N(x)$.

c) Considérons la fonction $N(x)$ sur l'ensemble compact $K = \{x \in \mathbb{R}^n\,;\ \|x\|_2 = 1\}$. Par le théorème 4.5, elle admet un minimum en un $u \in K$, c.-à-d. il existe $u \in K$ avec

$$N(u) \leq N(z) \qquad \text{pour tout} \quad z \in K.$$

Posons $C_1 := N(u)$; par (N1), on a $C_1 > 0$. Comme $x/\|x\|_2 \in K$ pour tout $x \in \mathbb{R}^n$ ($x \neq 0$), nous avons

$$C_1 = N(u) \leq N\Big(\frac{x}{\|x\|_2}\Big) = \frac{1}{\|x\|_2}\, N(x),$$

ce qui démontre la première inégalité de (4.4). □

I.5 Convergence uniforme et la courbe de Peano-Hilbert

Toutes les définitions et les résultats du cours "Analyse I" concernant la convergence uniforme d'une suite de fonctions se généralisent immédiatement à plusieurs dimensions. Rappelons qu'une suite de fonctions $f_k : A \to \mathbb{R}^m$, $A \subset \mathbb{R}^n$ *converge uniformement sur* A vers $f : A \to \mathbb{R}^m$, si

$$\forall \varepsilon > 0 \quad \exists N \geq 1 \quad \forall k \geq N \quad \forall x \in A \qquad \|f_k(x) - f(x)\| < \varepsilon.$$

Un exemple est le suivant.

Théorème 5.1 *Si une suite de fonctions continues $f_k : A \to \mathbb{R}^m$, $A \subset \mathbb{R}^n$ converge uniformément sur A vers une fonction $f : A \to \mathbb{R}^m$, cette fonction limite est continue.* □

La courbe de Peano-Hilbert L'image d'une fonction continue $\varphi : [0,1] \to \mathbb{R}^2$ est une courbe dans $\mathbb{R}^2$. Par exemple, la fonction $\varphi(t) = (t, 1-t)$ donne un segment de droite et la fonction $\varphi(t) = (\cos 2\pi t, \sin 2\pi t)$ un cercle. On peut se demander si l'image d'une fonction continue $\varphi : [0,1] \to \mathbb{R}^2$ peut remplir tout le carré $[0,1] \times [0,1]$. Peano (1890) et Hilbert (1891) ont découvert que ceci est possible: on divise chaque carré en quatre sous-carrés et on étiquette leurs centres récursivement, en suivant la direction de la courbe précédente (voir la figure I.10).

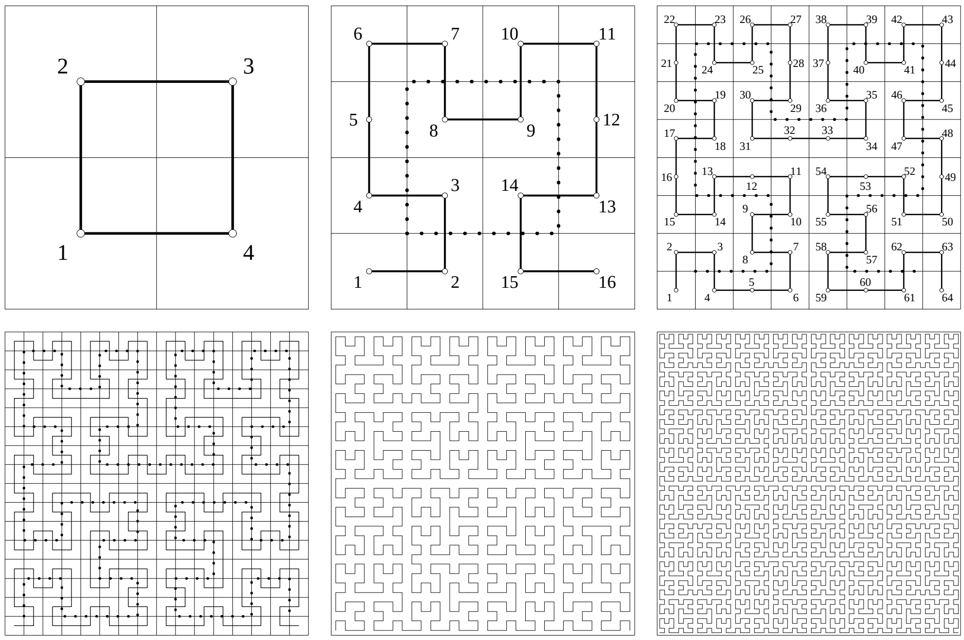


Figure I.10: La courbe de Peano-Hilbert

Une autre construction. Soit $\varphi(t) = (x(t), y(t)),\ \ 0 \le t \le 1$ une courbe continue arbitraire reliant les points $A = (0,0)$ pour $t = 0$ et $B = (1,0)$ pour $t = 1$ (voir la figure I.11). Nous définissons alors une nouvelle courbe $\Phi\varphi$ par

$$(\Phi\varphi)(t) = \begin{cases} \frac{1}{2}(y(4t), x(4t)) & \text{si } 0 \le t \le \frac{1}{4} \\ \frac{1}{2}(x(4t-1), 1 + y(4t-1)) & \text{si } \frac{1}{4} \le t \le \frac{2}{4} \\ \frac{1}{2}(1 + x(4t-2), 1 + y(4t-2)) & \text{si } \frac{2}{4} \le t \le \frac{3}{4} \\ \frac{1}{2}(2 - y(4t-3), 1 - x(4t-3)) & \text{si } \frac{3}{4} \le t \le 1. \end{cases}$$

Ainsi, nous avons à nouveau une courbe continue reliant $A = (0,0)$ pour $t = 0$ et $B = (1,0)$ pour $t = 1$ (voir le deuxième dessin de la figure I.11), et ainsi ce procédé peut être répété (troisième

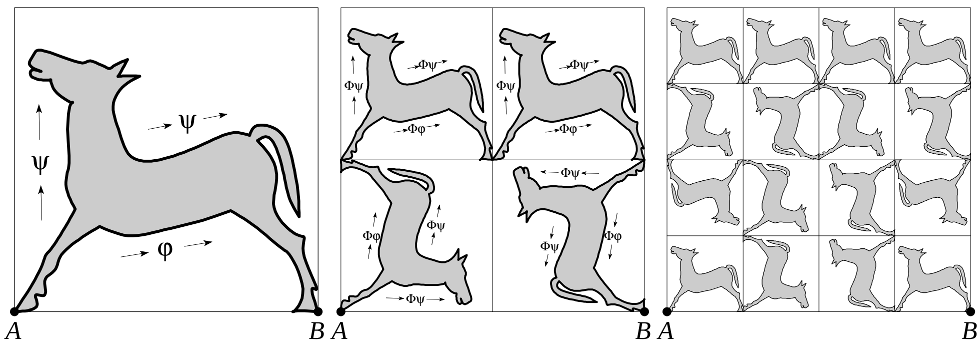


Figure I.11: Création de la courbe de Hilbert

dessin dans la figure I.11). Ceci définit une suite de fonctions $\varphi_0 = \varphi$, $\varphi_1 = \Phi\varphi_0$, $\varphi_2 = \Phi\varphi_1$, etc. Si nous commençons avec une autre courbe initiale $\psi(t)$ avec $\|\varphi(t)-\psi(t)\|_\infty \leq K$ pour $t \in [0,1]$, alors $\|\Phi\varphi(t) - \Phi\psi(t)\|_\infty \leq K/2$ (voir la figure I.11). Par conséquent,

$$\|\varphi_k(t) - \psi_k(t)\| \leq K \cdot 2^{-k}, \tag{5.1}$$

et, en prenant $\psi(t) = \varphi_m(t)$ et $K = 1$, on a

$$\|\varphi_k(t) - \varphi_{k+m}(t)\| \leq 2^{-k}. \tag{5.2}$$

Par (5.2), la suite $\varphi_k(t)$ converge uniformément (critère de Cauchy), et possède donc une limite continue $\varphi_\infty(t)$ (théorème 5.1). De plus, nous voyons avec (5.1) que la fonction limite est indépendante de la fonction initiale $\varphi_0(t)$.

Remarque. Il est intéressant de mentionner que les deux coordonnées $x(t)$ et $y(t)$ sont des exemples de fonctions continues, nulle part différentiables (sans démonstration).

I.6 Exercices

Chapter II

Calcul matriciel

Les fonctions linéaires sont des exemples très simples de fonctions à plusieurs variables, mais elles sont également très importantes. Nous verrons dans le chapitre III que beaucoup de fonctions (en fait toutes les fonctions différentiables) peuvent être localement approximées par des fonctions linéaires.

II.1 Rappel de l'algèbre linéaire

L'espace $\mathbb{R}^n$ est un espace vectoriel, c.-à-d. qu'on peut additionner ses éléments et on peut les multiplier par des scalaires $\lambda \in \mathbb{R}$.

Base canonique Chaque vecteur $x \in \mathbb{R}^n$ peut être écrit (de manière unique) comme

$$x = x_1 e_1 + x_2 e_2 + \ldots + x_n e_n \qquad \text{où} \qquad e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad e_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \quad e_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}. \tag{1.1}$$

On appelle $x_1,\, x_2, \ldots,\, x_n$ les *composantes* ou *coordonnées* du vecteur x et l'ensemble des vecteurs $\{e_1,\, e_2, \ldots,\, e_n\}$ la *base canonique* de $\mathbb{R}^n$. Dans ce chapitre nous écrivons les vecteurs de $\mathbb{R}^n$ comme des vecteurs colonnes.

Applications linéaires Une application linéaire est une fonction $f : \mathbb{R}^n \to \mathbb{R}^m$ satisfaisant

$$f(x+y) = f(x) + f(y), \quad f(\lambda x) = \lambda f(x) \qquad \text{pour tout} \quad x,\, y \in \mathbb{R}^n,\ \ \lambda \in \mathbb{R}. \tag{1.2}$$

En notant l'image de e_j par $f(e_j) = A_j = (a_{1j}, a_{2j}, \ldots, a_{mj})^t$, la linéarité de $f(x)$ implique que

$$f(x) = f(x_1 e_1 + \ldots + x_n e_n) = x_1 A_1 + \ldots + x_n A_n = Ax$$

où A est la matrice dont les colonnes sont $A_1, \ldots, A_n$. Celle-ci est une matrice $m \times n$ (m lignes et n colonnes). Le produit Ax est donné par

$$Ax = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + \ldots + a_{1n}x_n \\ \vdots \\ a_{m1}x_1 + \ldots + a_{mn}x_n \end{pmatrix}.$$

Chaque application linéaire peut donc être écrite sous la forme $f(x) = Ax$ où A est une matrice $m \times n$.

Bases arbitraires Un ensemble $\{S_1, S_2, \ldots, S_n\}$ de n vecteurs est une base de $\mathbb{R}^n$ si $S_1, S_2, \ldots, S_n$ sont linéairement indépendants. Chaque $x \in \mathbb{R}^n$ peut être écrit de manière unique comme

$$x = \xi_1 S_1 + \ldots + \xi_n S_n = S\,\xi \tag{1.3}$$

où S est la matrice $n \times n$ dont les colonnes sont les S_i, et ξ est le vecteur de composantes ξ_i (voir la figure II.1). On appelle alors $\xi_1, \xi_2, \ldots, \xi_n$ les coordonnées de x selon la base $\{S_1, S_2, \ldots, S_n\}$. Comme S est une matrice inversible, la relation (1.3) donne une bijection entre x et ξ.

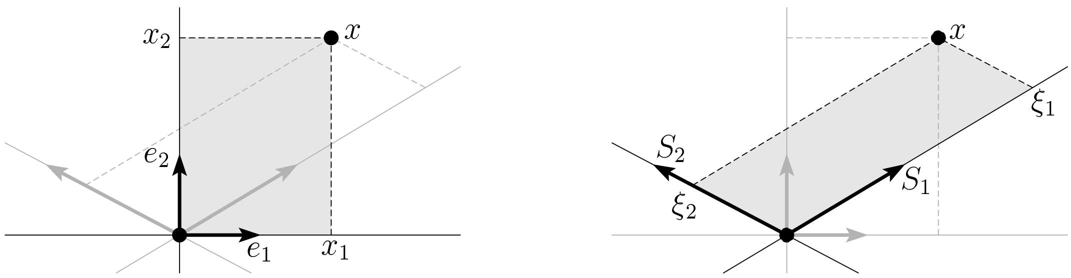


Figure II.1: Représentation de x dans la base canonique (gauche) et dans la base $\{S_1, S_2\}$ (droite)

Changement de coordonnées Considérons une application linéaire $f : \mathbb{R}^n \to \mathbb{R}^m$ donnée par $x \mapsto y = Ax$ où A est une matrice $m \times n$. Si l'on change les coordonnées dans l'espace de départ $\mathbb{R}^n$, ($x = S\xi$) et dans l'espace d'arrivé $\mathbb{R}^m$ ($y = T\eta$), l'application

$$y = Ax \qquad \text{devient} \qquad \eta = T^{-1}AS\,\xi.$$

Avec des bases $S = (S_1, \ldots, S_n)$ et $T = (T_1, \ldots, T_m)$ bien choisies, la matrice $T^{-1}AS$ peut devenir très simple. En particulier, on peut toujours trouver S et T tels que

$$T^{-1}AS = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}$$

où I_r designe la matrice unité de dimension r (voir le paragraphe III.7 du polycopié du cours "Algèbre I"). C'est-à-dire que dans les nouvelles coordonnées l'application $y = Ax$ devient simplement $\eta_i = \xi_i$ pour $i = 1, \ldots, r$ et $\eta_i = 0$ pour $i = r+1, \ldots, m$.

Si $f : \mathbb{R}^n \to \mathbb{R}^n$ est une application où les espaces de départ et d'arrivée sont identiques, on est interessé à faire le même changement de coordonnées pour x et y, i.e., $x = S\xi$ et $y = S\eta$. Dans cette situation, l'application devient $\eta = S^{-1}AS\,\xi$ et on cherche une transformation S telle que la matrice $S^{-1}AS$ soit plus simple que A.

Valeurs et vecteurs propres Pour transformer une matrice $n \times n$ en une forme simple, les valeurs propres et les vecteurs propres jouent un rôle important. On dit que $\lambda \in \mathbb{C}$ est une *valeur propre* de la matrice A s'il existe un vecteur $v \in \mathbb{C}^n$, $v \neq 0$ tel que

$$Av = \lambda v. \tag{1.4}$$

Un vecteur $v \neq 0$ satisfaisant (1.4) s'appelle *vecteur propre* de A correspondant à la valeur propre λ. Pour que λ soit une valeur propre de A, il faut que le système linéaire $(A - \lambda I)v = 0$ possède une solution non nulle. Par conséquent, λ doit nécessairement être un zéro du polynôme caractéristique

$$\chi_A(\lambda) := \det(A - \lambda I). \tag{1.5}$$

Matrices à coefficients complexes La considération des valeurs et vecteurs propres d'une matrice nécessite de travailler avec l'espace $\mathbb{C}^n$ des n-uples de nombres complexes. Celui-ci est également un espace vectoriel mais cette fois sur $\mathbb{C}$ (on peut additionner les éléments de $\mathbb{C}^n$ et on peut les multiplier par des scalaires $\lambda \in \mathbb{C}$).

La base canonique de $\mathbb{C}^n$ est la même que pour $\mathbb{R}^n$. Les applications linéaires $f : \mathbb{C}^n \to \mathbb{C}^m$ sont de nouveau de la forme $f(x) = Ax$ où cette fois les coefficients a_{ij} de la matrice A sont des nombres complexes. Une base de $\mathbb{C}^n$ est donnée par n vecteurs linéairement indépendants et un changement de coordonnées s'exprime de nouveau par les formules $x = S\xi$, $y = T\eta$ tel que $y = Ax$ devient $\eta = T^{-1}AS\,\xi$. Ici, toutes les matrices sont à coefficients complexes. La définition de valeurs et vecteurs propres ne change pas par rapport à $\mathbb{R}^n$.

Pour un nombre complexe $\lambda = \alpha + i\beta$, son conjugé est donné par $\overline{\lambda} = \alpha - i\beta$ et le carré de sa valeur absolue est $|\lambda|^2 = \overline{\lambda} \cdot \lambda = \alpha^2 + \beta^2$. Pour un vecteur $x \in \mathbb{C}^n$ et pour une matrice A à coefficients complexes, on dénote par $\overline{x}$ et $\overline{A}$ les objets où tous les éléments sont remplacés par leur conjugé. Le vecteur transposé conjugé $\overline{x}^t$ est noté x^* et, similairement, la matrice transposée conjugée de A est $A^* := \overline{A}^t$. Elle s'appelle aussi matrice adjointe de A.

Si l'on identifie $\mathbb{C}$ avec $\mathbb{R}^2$ ($\alpha + i\beta \leftrightarrow (\alpha, \beta)$), la valeur absolue du nombre complexe $\alpha + i\beta$ correspond à la norme euclidienne du vecteur (α, β). En identifiant $\mathbb{C}^n$ avec $\mathbb{R}^{2n}$, on peut introduire une norme sur $\mathbb{C}^n$

$$\|z\| = \sqrt{|z_1|^2 + \ldots + |z_n|^2} = \sqrt{x_1^2 + y_1^2 + \ldots + x_n^2 + y_n^2} \tag{1.6}$$

si $z = (z_1, \ldots, z_n)^t$ et $z_j = x_j + iy_j$. Les propriétés (N1), (N2) et (N3) du théorème I.1.1 restent vraies et on a $\|\lambda x\| = |\lambda| \cdot \|x\|$ pour tout $x \in \mathbb{C}^n$ et $\lambda \in \mathbb{C}$. En définissant le produit scalaire de deux vecteurs $x,\ y \in \mathbb{C}^n$ par

$$\langle x, y \rangle := x^* y = \sum_{j=1}^{n} \overline{x}_j y_j, \tag{1.7}$$

on obtient $\|x\| = \sqrt{\langle x, x \rangle}$ et l'inégalité de Cauchy-Schwarz (I.1.5) reste vraie dans $\mathbb{C}^n$.

II.2 Forme normale de Schur

Considérons une application linéaire $y = Ax$ où A est une matrice carrée de dimension n à coefficients réels ou complexes. Le but de ce paragraphe est de simplifier A en utilisant un changement de coordonnées $y = S\eta$, $x = S\xi$ (deux fois la même matrice S) qui laisse la distance entre deux vecteurs invariante (isométrie). Ceci signifie qu'on demande $d(Sx, Sz) = d(x, z)$ ou $\|Sx - Sz\| = \|x - z\|$ ou $(x - z)^* S^* S(x - z) = (x - z)^*(x - z)$.

Définition 2.1 (matrices orthogonales et unitaires) *Une base $\{S_1, \ldots, S_n\}$ de $\mathbb{R}^n$ ou de $\mathbb{C}^n$ est une base orthonormale si $\|S_j\| = 1$ pour tout j et si $\langle S_j, S_k \rangle = 0$ pour $j \neq k$.*

Une matrice carrée S à coefficients réels s'appelle orthogonale si $S^t S = I$.

Une matrice carrée S à coefficients complexes s'appelle unitaire si $S^* S = I$.

Ces transformations sont extrêmement importantes pour un calcul sur ordinateur parce qu'elles n'amplifient pas les erreurs d'arrondi (voir le cours “Analyse Numérique”, 2ème année).

Remarquons encore que le produit de matrices orthogonales (resp. unitaires) est orthogonal (resp. unitaire) et qu'on a $S^{-1} = S^t$ pour des matrices orthogonales et $S^{-1} = S^*$ pour des matrices unitaires.

Exemple 2.2 Toutes le matrices orthogonales de dimension 2 sont données par

$$U_1 = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \qquad \text{et} \qquad U_2 = \begin{pmatrix} \cos\alpha & \sin\alpha \\ \sin\alpha & -\cos\alpha \end{pmatrix}. \tag{2.1}$$

La transformation avec la matrice U_1 correspond à une rotation d'angle α (observer $\det U_1 = 1$), celle avec U_2 est une réflexion suivi d'une rotation ($\det U_2 = -1$).

Des exemples de matrices orthogonales de dimension 3 sont

$$R_1(\alpha)=\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha \\ 0 & \sin\alpha & \cos\alpha \end{pmatrix} \quad R_2(\beta)=\begin{pmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{pmatrix} \quad R_3(\gamma)=\begin{pmatrix} \cos\gamma & -\sin\gamma & 0 \\ \sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Chaque matrice orthogonale de dimension 3 avec déterminant égal à $+1$ peut être écrite sous la forme $U = R_3(\varphi)R_2(\theta)R_3(\psi)$ (sans démonstration). Les trois angles φ, θ, ψ s'appellent les "angles d'Euler" de la rotation dans $\mathbb{R}^3$.

Théorème 2.3 (forme normale de Schur, version complexe) *Soit A une matrice $n \times n$ à coefficients réels ou complexes. Il existe une matrice unitaire U (i.e., $U^*U = I$) telle que*

$$U^{-1}A\,U = U^*A\,U = \begin{pmatrix} \lambda_1 & * & \cdots & * \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix} \tag{2.2}$$

est triangulaire avec les valeurs propres $\lambda_1, \ldots, \lambda_n$ de A sur la diagonale.

Démonstration. Soit λ_1 un zéro du polynôme caractéristique (1.5) et v_1 un vecteur propre correspondant. Choisissons une matrice unitaire U_1 dont la première colonne est v_1 (procédé d'orthonormalisation de Gram-Schmidt, voir le paragraphe VI.1 du polycopié du cours "Algèbre I"). Avec cette matrice on obtient

$$A\,U_1 = U_1 \begin{pmatrix} \lambda_1 & s^t \\ 0 & B \end{pmatrix} \qquad \text{i.e.} \qquad U_1^*A\,U_1 = \begin{pmatrix} \lambda_1 & s^t \\ 0 & B \end{pmatrix}$$

où B est une matrice $(n-1) \times (n-1)$. Supposons, par récurrence sur la dimension de la matrice, qu'il existe une matrice unitaire $\widehat{U}$ telle que $\widehat{U}^*B\,\widehat{U}$ est sous forme triangulaire. Alors, avec la matrice unitaire

$$U := U_1 \begin{pmatrix} 1 & 0 \\ 0 & \widehat{U} \end{pmatrix}$$

on obtient

$$U^*A\,U = \begin{pmatrix} 1 & 0 \\ 0 & \widehat{U}^* \end{pmatrix} \begin{pmatrix} \lambda_1 & s^t \\ 0 & B \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \widehat{U} \end{pmatrix} = \begin{pmatrix} \lambda_1 & s^t\widehat{U} \\ 0 & \widehat{U}^*B\,\widehat{U} \end{pmatrix}$$

qui est sous la forme souhaitée. □

On appelle une matrice A *normale* si $A^*A = AA^*$. Pour une matrice à coefficients réels ceci signifie que $A^tA = AA^t$. Des exemples de matrices normales réelles sont les matrices symétriques ($A^t = A$) ou anti-symétriques ($A^t = -A$). Pour des matrices normales, le théorème précédent donne une matrice diagonale.

Corollaire 2.4 *Soit A une matrice normale (i.e., $A^*A = AA^*$). Il existe alors une matrice unitaire U telle que*

$$U^{-1}A\,U = U^*A\,U = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_n).$$

Démonstration. Soit U la matrice unitaire obtenue par le théorème 2.3 et soit S la matrice triangulaire de (2.2). Comme A est normale, on a

$$S^*S = (U^*A\,U)^*(U^*A\,U) = U^*A^*UU^*A\,U = U^*A^*A\,U = U^*AA^*U = \ldots = SS^*,$$

c.-à-d. S est aussi normale. Cette condition implique que la norme de la $j$ème colonne de S est égal à la norme de sa $j$ème ligne. Pour $j = 1$, ceci implique que tous les éléments de la première ligne de (2.2) sont nuls sauf celui sur la diagonale. Ensuite on considère $j = 2$ et on déduit de la même manière que les éléments de la deuxième ligne en dehors de la diagonale sont nuls. En continuant ce raisonnement, on obtient le résultat. □

Traitons maintenant le cas où A est une matrice à coefficients réels. Il est impossible de trouver sans restriction une décomposition de la forme (2.2) à l'aide d'une matrice orthogonale U. Ceci impliquerait que les λ_i soient réels et donne ainsi une contradiction.

Théorème 2.5 (forme normale de Schur, version réelle) *Soit A une matrice $n \times n$ à coefficients réels. Il existe une matrice orthogonale U (i.e., $U^tU = I$) telle que*

$$U^{-1}A\,U = U^tA\,U = \begin{pmatrix} \Lambda_1 & * & \cdots & * \\ 0 & \Lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \cdots & 0 & \Lambda_n \end{pmatrix} \tag{2.3}$$

où Λ_j est soit la matrice (λ_j) de dimension 1, soit la matrice $\begin{pmatrix} \alpha_j & \beta_j/\nu_j \\ -\nu_j\beta_j & \alpha_j \end{pmatrix}$ de dimension 2. Les valeurs propres de A sont λ_j et $\alpha_j \pm i\beta_j$.

Démonstration. Nous suivons la démonstration du théorème 2.3.

Si λ_1 est une valeur propre réelle de A, on peut choisir le vecteur propre v_1 tel qu'il soit réel et de norme 1. En construisant une matrice orthogonale U_1 dont la première colonne est v_1, la première colonne de $U_1^tA\,U_1$ est sous la forme souhaitée.

Soit alors $\lambda_1 = \alpha + i\beta$ avec $\beta \neq 0$ et $v_1 = u + iw$. On peut supposer $u^tw = 0$ car, sinon, on considère $(1+i\mu)(u+iw) = (u-\mu w)+i(\mu u+w)$ qui est aussi un vecteur propre et on détermine μ afin que $(u-\mu w) \perp (\mu u+w)$. Ceci donne l'équation quadratique $\mu^2u^tw+\mu(\|w\|-\|u\|)-u^tw = 0$ qui possède toujours une solution réelle. Avec $u_1 := u/\|u\|$, $w_1 := w/\|w\|$ et $\nu := \|w\|/\|u\|$, $u_1 + i\nu w_1$ est un vecteur propre de A, c.-à-d.

$$A(u_1 + i\nu w_1) = (\alpha + i\beta)(u_1 + i\nu w_1)$$

et en séparant les parties réelles et imaginaires on obtient

$$A(u_1, w_1) = (u_1, w_1)\begin{pmatrix} \alpha & \beta/\nu \\ -\nu\beta & \alpha \end{pmatrix}.$$

On complète les deux vecteurs u_1, w_1 en une base orthonormale (pour obtenir U_1) et on voit que les deux premières colonnes de $U_1^tA\,U_1$ sont comme demandées dans (2.3).

On termine la démonstration par induction sur la dimension de la matrice exactement comme dans la démonstration du théorème 2.3. □

Corollaire 2.6 *Si A est une matrice réelle et symétrique (i.e., $A^t = A$), il existe une matrice orthogonale U telle que*

$$U^{-1}A\,U = U^tA\,U = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_n). \tag{2.4}$$

Les valeurs propres λ_j de A sont réelles.

Démonstration. Si A est symétrique, la matrice $U^tA\,U$ de (2.3) est aussi symétrique. Ceci implique que les Λ_j sont de dimension 1. Par conséquent, toutes les valeurs propres sont réelles et $U^tA\,U$ est diagonale. □

Paradoxe Les démonstrations précédentes nécessitent la connaissance des valeurs propres pour trouver la forme normale de Schur. Au cours "Analyse Numérique", nous apprendrons l'algorithme QR qui est un procédé itératif pour calculer la forme normale de Schur (sans utiliser explicitement les valeurs propres). Ceci est alors un algorithme (en fait le plus important) pour calculer les valeurs propres d'une matrice.

II.3 Formes quadratiques

Considérons une fonction quadratique à deux variables

$$f(x_1, x_2) = ax_1^2 + 2bx_1x_2 + cx_2^2 + dx_1 + ex_2 + g \tag{3.1}$$

et étudions l'ensemble $\mathcal{M} = \{(x_1, x_2)\,;\, f(x_1, x_2) = 0\}$ qui constitue une courbe dans $\mathbb{R}^2$. Elle s'appelle *conique* (comme intersection d'un cône avec un plan). Par exemple, la fonction

$$f_1(x_1, x_2) = x_1^2 + 2bx_1x_2 + x_2^2 - 5x_1 - 4x_2 + g \tag{3.2}$$

donne les coniques de la figure II.2. Comment peut-on déterminer s'il s'agit d'une ellipse, d'une hyperbole ou d'une parabole ?

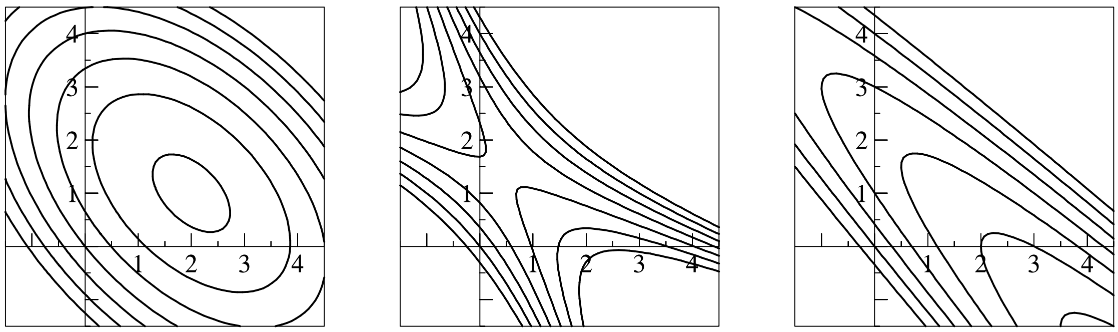


Figure II.2: Les coniques correspondantes à la fonction (3.2) avec $b = 0.5$ (gauche), $b = 1.5$ (milieu), $b = 1$ (droite) et avec plusieurs valeurs de g

Nous écrivons la fonction (3.1) sous forme matricielle

$$f(x) = x^tA\,x + B^tx + C \tag{3.3}$$

où $x = (x_1, x_2)^t$ et

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}, \quad B = \begin{pmatrix} d \\ e \end{pmatrix}, \quad C = g,$$

et nous essayons de transformer cette fonction par un changement de coordonnées en une fonction plus simple.

Considérons tout de suite une fonction (3.3) dans $\mathbb{R}^n$ où A est une matrice symétrique de dimension n, $B \in \mathbb{R}^n$ et $C \in \mathbb{R}$. Pour simplifier une telle fonction, nous procédons en deux pas: une rotation et une translation.

Rotation Comme A est symétrique, il existe une matrice orthogonale U (i.e. $U^tU = I$) avec $\det U = 1$ (corollaire 2.6) telle que

$$U^t A\, U = D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n).$$

Avec le changement de variables $x = Uy$, on obtient alors (avec $E^t = B^tU$)

$$x^t A\, x + B^t x + C = y^t D\, y + E^t y + C.$$

Translation Avec la translation $y = z + c$ (c est un vecteur de $\mathbb{R}^n$), la forme quadratique devient

$$y^t D\, y + E^t y + C = z^t D\, z + (2D\, c + E)^t z + c^t D\, c + E^t c + C.$$

Si $\lambda_i \neq 0$, on définit $c_i = -d_i/(2\lambda_i)$ afin d'annuler le terme linéaire en z_i. Finalement, si un des λ_i vaut zéro, par exemple $\lambda_n = 0$, mais si $d_n \neq 0$, on peut choisir c_n afin d'annuler le terme constant.

Avec toutes ces simplifications et en notant la variable de nouveau par $x = (x_1, \ldots, x_n)^t$, on obtient

$$\mathcal{M} = \{x \in \mathbb{R}^n \mid x^t D\, x + B^t x + C = 0\},$$

où $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ est une matrice diagonale, $D\, B = 0$, et $C = 0$ si $B \neq 0$. On distingue alors les cas suivants:

- *hyperellipsoïde*: si $\lambda_i > 0$ pour $i = 1, \ldots, n$ (ou tous les λ_i négatifs)
- *hyperhyperboloïde*: s'il y a au moins un λ_i positif, et au moins un λ_i négatif
- *hyperparaboloïde*: si $\lambda_i > 0$ pour $i = 1, \ldots, n-1$, $\lambda_n = 0$, et $d_n \neq 0$
- *hypersphère*: si $\lambda_i = 1$ pour $i = 1, \ldots, n$
- *hyperplan*: si $\lambda_i = 0$ pour $i = 1, \ldots, n$.

Cas $\mathbf{n} = 2$ Selon les signes des deux valeurs propres on obtient (avec x, y au lieu de x_1, x_2)

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - 1 = 0 \ \text{(ellipse)}, \quad \frac{x^2}{a^2} - \frac{y^2}{b^2} \pm 1 = 0 \ \text{(hyperbole)}, \quad x^2 - 2py = 0 \ \text{(parabole)}.$$

Cas $\mathbf{n} = 3$ En négligeant les cas dégénérés où $\mathcal{M}$ est l'ensemble vide, un point, ou une surface cylindrique (c.-à-d. $x^t D\, x + B^t x + C$ ne dépend que de deux variables), on a les surfaces suivantes (nous écrivons x, y, z au lieu de x_1, x_2, x_3):

(A) $\displaystyle \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} - 1 = 0$ (ellipsoïde)

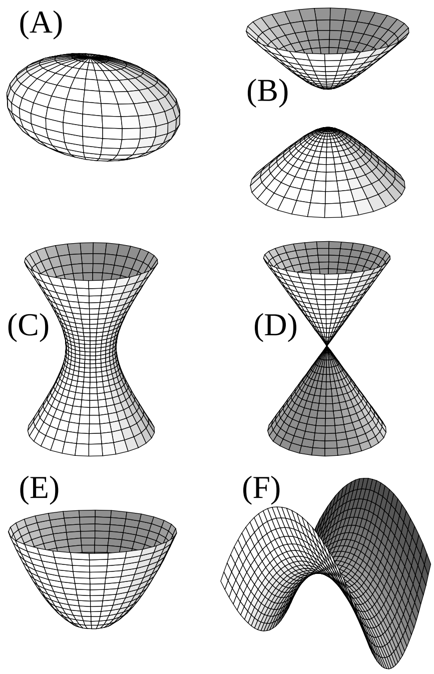
(B) $\displaystyle \frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} + 1 = 0$ (hyperboloïde à deux nappes)

(C) $\displaystyle \frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} - 1 = 0$ (hyperboloïde à une nappe)

(D) $\displaystyle \frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 0$ (cône)

(E) $\displaystyle \frac{x^2}{a^2} + \frac{y^2}{b^2} - 2pz = 0$ (paraboloïde elliptique)

(F) $\displaystyle \frac{x^2}{a^2} - \frac{y^2}{b^2} - 2pz = 0$ (paraboloïde hyperbolique)



II.4 Matrices définies positives

Dans ce paragraphe nous étudions des matrices symétriques pour lesquelles la forme quadratique $f(x) = x^t A\, x$ est toujours non-négative. Dans les applications, x est souvent un vecteur de vitesse, A une matrice de masse et $x^t A\, x$ l'énergie cinétique.

Définition 4.1 (matrice définie positive) *Une matrice symétrique A s'appelle définie positive si*

$$x^t A\, x > 0 \qquad \textit{pour tout} \;\; x \in \mathbb{R}^n,\; x \neq 0. \tag{4.1}$$

Elle s'appelle semi-définie positive si

$$x^t A\, x \geq 0 \qquad \textit{pour tout} \;\; x \in \mathbb{R}^n. \tag{4.2}$$

Remarquons qu'une matrice définie positive nous permet de définir une norme sur $\mathbb{R}^n$,

$$\|x\|_A := \sqrt{x^t A\, x}\;. \tag{4.3}$$

Motivé par des applications, cette norme s'appelle souvent *norme d'énergie*. Les propriétés (N1) et (N2) d'une norme sont facile à vérifier. L'inégalité du triangle (N3) se démontre comme dans la preuve du théorème I.1.1 à l'aide de l'inégalité de Cauchy-Schwarz

$$|x^t A\, y| \leq \|x\|_A \cdot \|y\|_A. \tag{4.4}$$

Cette dernière suit de $0 \leq (x + \mu y)^t A(x + \mu y) = x^t A\, x + 2\mu x^t A\, y + \mu^2 y^t A\, y$ en posant $\mu = -x^t A\, y / y^t A\, y$ (comme on l'a fait pour la norme euclidienne).

Théorème 4.2 *Nous avons les critères suivants (λ_i sont les valeurs propres de A):*

$$\begin{aligned} A \textit{ est définie positive} \quad &\Longleftrightarrow \quad \lambda_i > 0 \quad \textit{pour} \;\; i = 1, \ldots, n, \\ A \textit{ est semi-définie positive} \quad &\Longleftrightarrow \quad \lambda_i \geq 0 \quad \textit{pour} \;\; i = 1, \ldots, n. \end{aligned}$$

Démonstration. On utilise le corollaire 2.6 qui garantit l'existence d'une matrice orthogonale telle que $U^t A\, U = D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Pour $x \in \mathbb{R}^n$ arbitraire, définissons $y \in \mathbb{R}^n$ par $x = Uy$. Nous avons alors

$$x^t A\, x = y^t D\, y = \sum_{j=1}^{n} \lambda_j y_j^2 \tag{4.5}$$

et on voit que (4.1) est satisfait pour tout $x \neq 0$ (i.e. l'expression (4.5) est positive pour tout $y \neq 0$) si et seulement si tous les λ_i sont positifs.

La démonstration pour le critère d'une matrice semi-définie positive est analogue. □

La caractérisation des matrices définies positives du théorème précédent nécessite la connaissance de toutes les valeurs propres de la matrice. Il est connu qu'il n'y a pas d'algorithme fini qui permet de faire ce calcul pour toutes les matrices.

Le théorème suivant nous permet de décider si une matrice symétrique A est définie positive ou non avec un nombre fini d'opérations.

Théorème 4.3 *Pour une matrice symétrique A de dimension n nous considérons les sous-matrices*

$$A_1 = (a_{11}), \quad A_2 = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \quad A_3 = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}, \quad \ldots\;, \quad A_n = A.$$

Alors,

$$A \textit{ est définie positive} \quad \Longleftrightarrow \quad \det A_i > 0 \quad \textit{pour} \;\; i = 1, \ldots, n.$$

Démonstration. Nécessité. Le déterminant d'une matrice définie positive A est positif car $\det A = \lambda_1 \cdot \ldots \cdot \lambda_n > 0$. Pour un vecteur $x = (x_1, \ldots, x_n)^t \in \mathbb{R}^n$, notons par $P_j x = (x_1, \ldots, x_j, 0, \ldots, 0)^t$ sa projection sur les j premières composantes et par $X_j = (x_1, \ldots, x_j)^t$. On a alors

$$X_j^t A_j X_j = (P_j x)^t A\,(P_j x) > 0 \qquad \text{pour tout } \; X_j \neq 0,$$

ce qui implique que A_j est définie positive et donc $\det A_j > 0$.

Suffisance. Cette partie de la démonstration se fait par récurrence sur la dimension n de la matrice. Pour $n = 1$, il n'y a rien à démontrer. Supposons alors que A_{n-1} soit définie positive et que $\det A_n > 0$. On peut donc diagonaliser A_{n-1} à l'aide d'une matrice orthogonale U. Ceci donne

$$\begin{pmatrix} U^t & 0 \\ 0 & 1 \end{pmatrix} A_n \begin{pmatrix} U & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} D & b \\ b^t & c \end{pmatrix} \tag{4.6}$$

où $D = \operatorname{diag}(\mu_1, \ldots, \mu_{n-1})$ avec $\mu_j > 0$ (les valeurs propres de A_{n-1}), $b = (b_1, \ldots, b_{n-1})^t \in \mathbb{R}^{n-1}$ et $c \in \mathbb{R}$. En observant que le déterminant de la matrice (4.6) est égal à $\det A_n$ et en développant le déterminant de (4.6) relativement à la dernière ligne (voir le paragraphe IV.3 du polycopié de “Algèbre I”), on obtient

$$\det A_n = \mu_1 \cdot \ldots \cdot \mu_{n-1} \Big(-\frac{b_1^2}{\mu_1} - \ldots - \frac{b_{n-1}^2}{\mu_{n-1}} + c \Big) > 0\,.$$

D'autre part, on a pour $x \in \mathbb{R}^n$ et avec $y \in \mathbb{R}^n$ donné par $x = \begin{pmatrix} U & 0 \\ 0 & 1 \end{pmatrix} y$ que

$$x^t A_n x = y^t \begin{pmatrix} D & b \\ b^t & c \end{pmatrix} y = \sum_{j=1}^{n-1} \mu_j y_j^2 + 2 y_n \sum_{j=1}^{n-1} b_j y_j + c y_n^2. \tag{4.7}$$

L'inégalité de Cauchy-Schwarz (I.1.5) appliquée aux vecteurs $(b_j / \sqrt{\mu_j})_j$ et $(\sqrt{\mu_j}\, y_j)_j$ donne

$$\Big| \sum_{j=1}^{n-1} b_j y_j \Big|^2 \leq \Big(\sum_{j=1}^{n-1} \frac{b_j^2}{\mu_j} \Big) \Big(\sum_{j=1}^{n-1} \mu_j y_j^2 \Big) \leq \; c \sum_{j=1}^{n-1} \mu_j y_j^2\,.$$

En insérant cette estimation dans (4.7), nous obtenons $x^t A_n x \geq \Big(\sqrt{\sum_j \mu_j y_j^2} - \sqrt{c y_n^2} \Big)^2 \geq 0$ et, par conséquent, $\lambda_j \geq 0$ pour les valeurs propres de A_n. Comme $\det A_n > 0$, ces valeurs propres sont nécessairement strictement positives. □

II.5 Norme d'une matrice

Le but de ce paragraphe est de majorer l'expression Ax où A est une matrice $m \times n$.

Définition 5.1 (norme d'une matrice) *Soit A une matrice $m \times n$. On définit*

$$\|A\| = \sup\{\, \|Ax\|\;;\; \|x\| \leq 1 \,\} \tag{5.1}$$

et on appelle ce nombre la norme de la matrice A ou la norme de l'application linéaire $f(x) = Ax$.

Comme la boule unité $K := \{x \in \mathbb{R}^n;\, \|x\| \leq 1\}$ est un ensemble compact et la fonction linéaire $f(x) = Ax$ est partout continue, le théorème I.4.5 implique que $\|A\|$ est finie et qu'on pourrait remplacer le “$\sup$” par un “$\max$”.

Des expressions équivalentes pour $\|A\|$ sont

$$\|A\| = \sup\{\, \|Ax\| \;;\; \|x\| = 1 \,\} = \sup\Big\{ \frac{\|Ax\|}{\|x\|} \;;\; x \neq 0 \Big\}. \tag{5.2}$$

De cette dernière formule, on voit que $\|A\|$ est le plus petit nombre réel tel que

$$\|Ax\| \leq \|A\| \cdot \|x\| \qquad \text{pour tout } \; x \in \mathbb{R}^n. \tag{5.3}$$

L'inégalité (5.3) est fondamentale pour tous les calculs avec des applications linéaires.

Remarquons encore que $\|A\|$ dépend des normes choisies dans les ensembles de départ et d'arrivée. Si l'on choisit la norme $\|\cdot\|_p$ ($p = 1, 2$ ou ∞) dans les deux espaces, on dénote la norme de A par $\|A\|_p = \sup\{\, \|Ax\|_p \;;\; \|x\|_p \leq 1 \,\}$. Considérons par exemple la matrice

$$A = \begin{pmatrix} 1.4 & 1.0 \\ 0.4 & 1.7 \end{pmatrix}.$$

La figure II.3 montre l'image $f(B)$ de la boule unité B pour les trois normes (nous avons aussi dessiné les images de quelques droites verticales et horizontales). En traitillé nous avons ajouté la boule unité avec rayon $\|A\|_p$. Elle possède le plus petit rayon parmi les boules contenant l'image $f(B)$. Pour cette matrice on trouve que $\|A\|_1 = 2.7$, $\|A\|_2 = 2.29466$ et $\|A\|_\infty = 2.4$ (voir aussi le théorème 5.3 plus bas).

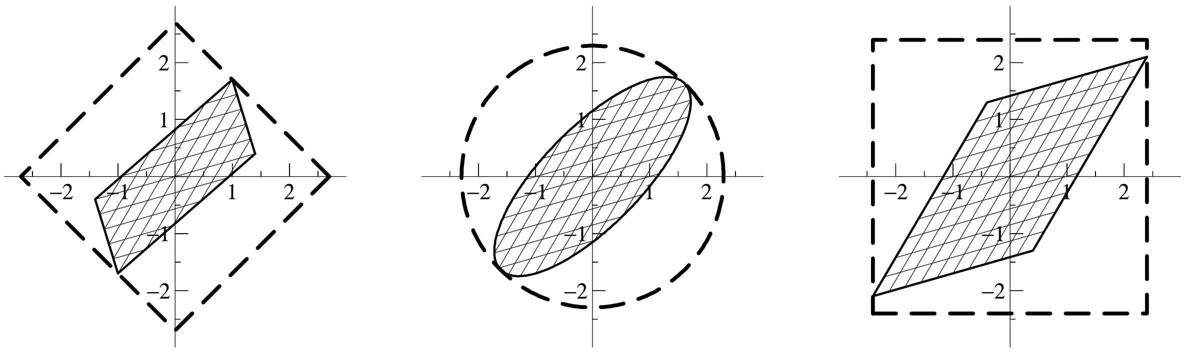


Figure II.3: Norme d'une matrice: ℓ_1 (gauche), euclidienne (milieu), maximum (droite)

Pour justifier la notation “norme d'une matrice” démontrons le résultat suivant.

Lemme 5.2 *L'application $A \mapsto \|A\|$ (de l'ensemble des matrices $M_{m,n}(\mathbb{R})$ dans $\mathbb{R}$) est une norme, c.-à-d. elle satisfait les propriétés (N1), (N2) et (N3). De plus on a*

$$\|I\| = 1 \qquad \textit{et} \qquad \|AB\| \leq \|A\| \cdot \|B\| \tag{5.4}$$

où I est la matrice identité et A, B sont deux matrices dont le produit AB est bien défini. Pour avoir $\|I\| = 1$ il faut choisir la même norme dans les ensembles de départ et d'arrivée.

Démonstration. Les propriétés (N1) et (N2) d'une norme sont faciles à vérifier. Démontrons (N3): pour $A, B \in M_{m,n}(\mathbb{R})$ nous avons

$$\|(A+B)x\| \leq \|Ax\| + \|Bx\| \leq (\|A\| + \|B\|)\|x\|.$$

En divisant cette relation par $\|x\|$ et en prenant le supremum, nous obtenons l'inégalité du triangle $\|A+B\| \leq \|A\| + \|B\|$.

La propriété $\|I\| = 1$ est évidente. Pour démontrer l'estimation de $\|AB\|$, nous appliquons deux fois l'inégalité fondamentale (5.3),

$$\|(AB)x\| \leq \|A\| \cdot \|Bx\| \leq \|A\| \cdot \|B\| \cdot \|x\|.$$

Ensuite, nous divisons cette relation par $\|x\|$ et nous prenons le supremum sur $x \neq 0$. □

Pour la norme ℓ_1, la norme euclidienne et la norme maximum, il existe une formule explicite pour $\|A\|_p = \sup\{\, \|Ax\|_p \, ; \; \|x\|_p \le 1 \,\}$.

Théorème 5.3 *Pour une matrice A de type $m \times n$, on a les formules*

$$\|A\|_1 \;=\; \max_{j=1,\dots,n}\Big(\sum_{i=1}^m |a_{ij}|\Big), \qquad \|A\|_\infty \;=\; \max_{i=1,\dots,m}\Big(\sum_{j=1}^n |a_{ij}|\Big),$$
$$\|A\|_2 \;=\; \sqrt{\textit{plus grande valeur propre de } A^tA}.$$

Démonstration. Pour la norme $\|x\|_1$, on a

$$\|Ax\|_1 = \sum_{i=1}^m \Big|\sum_{j=1}^n a_{ij}x_j\Big| \le \sum_{i=1}^m\sum_{j=1}^n |a_{ij}|\cdot|x_j| = \sum_{j=1}^n \Big(\sum_{i=1}^m |a_{ij}|\Big)|x_j| \le \max_{j=1,\dots,n}\Big(\sum_{i=1}^m |a_{ij}|\Big)\cdot\|x\|_1 .$$

On en déduit que $\|A\|_1 \le \max_j(\sum_i |a_{ij}|)$. Pour montrer l'égalité, on choisit un j_0 avec $\max_j(\sum_i |a_{ij}|) = \sum_i |a_{ij_0}|$ et on pose $x = (0,\dots,0,1,0,\dots,0)^T$, où 1 est à la position j_0. Ce choix de x donne l'égalité dans l'estimation ci-dessus, ce qui démontre que $\|A\|_1$ ne peut pas être plus petite que $\max_j(\sum_i |a_{ij}|)$. La formule pour la norme $\|x\|_\infty$ se démontre de la même manière.

La matrice A^tA étant symétrique et semi-définie positive ($x^tA^tAx = \|Ax\|_2^2 \ge 0$), il existe une matrice orthogonale U ($U^tU = I$) telle que $U^tA^tA\,U = \mathrm{diag}(\lambda_1,\dots,\lambda_n)$, où $\lambda_i \ge 0$ sont les valeurs propres de A^tA. On obtient avec la transformation $x = Uy$ et en utilisant $\|x\|_2 = \|y\|_2$ que

$$\|Ax\|_2^2 = x^tA^tAx = y^tU^tA^tA\,Uy = \sum_{i=1}^n \lambda_i|y_i|^2 \le \lambda_{\max}\,\|y\|_2^2 = \lambda_{\max}\,\|x\|_2^2 .$$

Ceci implique que $\|A\|_2 \le \sqrt{\lambda_{\max}}$. Pour montrer l'égalité, on pose x égal au vecteur propre de A^tA qui correspond à la valeur propre $\lambda_{\max}$. □

II.6 Applications bilinéaires et multilinéaires

En vue d'une étude des dérivées supérieures d'une fonction à plusieurs variables (chapitre III), nous préparons quelques notations concernant des applications bilinéaires et multilinéaires.

Définition 6.1 (application bilinéaire) *Une application $B : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^m$ s'appelle bilinéaire si elle est linéaire dans chacun des arguments, c.-à-d. si les applications suivantes sont linéaires:*

$$x \mapsto B(x,y) \quad \textit{pour tout } y \in \mathbb{R}^p \textit{ et} \qquad y \mapsto B(x,y) \quad \textit{pour tout } x \in \mathbb{R}^n.$$

Si l'on écrit les vecteurs x et y dans la base canonique, la bilinéarité implique que

$$B(x,y) = B\Big(\sum_{j=1}^n x_je_j\,,\,\sum_{k=1}^m y_ke_k\Big) = \sum_{j=1}^n\sum_{k=1}^m x_jy_kB(e_j,e_k)$$

Chaque application bilinéaire peut donc être écrite sous la forme

$$B(x,y) = \Big(\sum_{j=1}^n\sum_{k=1}^p b^i_{jk}x_jy_k\Big)_{i=1}^m . \tag{6.1}$$

Pour $m = 1$, elle devient simplement

$$B(x,y) = x^tBy \tag{6.2}$$

avec une matrice $B = (b_{jk})_{j,k}$ du type $n \times p$. Pour le cas général, la $i$ème composante de $B(x,y)$ est de la forme suivante : $B_i(x,y) = x^tB_iy$ avec $(b^i_{jk})_{j,k}$.

Norme d'une application bilinéaire Comme pour des applications linéaires, on définit

$$\|B\| = \sup_{\|x\|\leq 1,\, \|y\|\leq 1} \|B(x,y)\| = \sup_{x\neq 0,\, y\neq 0} \frac{\|B(x,y)\|}{\|x\|\cdot\|y\|} \tag{6.3}$$

et on obtient que $\|B\|$ est le plus petit nombre satisfaisant

$$\|B(x,y)\| \leq \|B\|\cdot\|x\|\cdot\|y\| \qquad \text{pour tout} \quad x\in\mathbb{R}^n,\, y\in\mathbb{R}^p. \tag{6.4}$$

L'expression $\|B\|$ de (6.3) dépend évidemment des normes choisies dans les trois espaces $\mathbb{R}^n$, $\mathbb{R}^p$ et $\mathbb{R}^m$. Si l'on prend la même norme $\|\cdot\|_p$ (avec $p=1$, 2 ou ∞) dans ces trois espaces, on dénote la norme de l'application bilinéaire par $\|B\|_p$.

Théorème 6.2 *On a les majorations*

$$\|B\|_1 \leq \sum_{i=1}^m \max_{j,k} |b^i_{jk}|, \qquad \|B\|_2 \leq \sqrt{\sum_{i=1}^m\sum_{j=1}^n\sum_{k=1}^p |b^i_{jk}|^2}, \qquad \|B\|_\infty \leq \max_{i=1,\ldots,m}\sum_{j=1}^n\sum_{k=1}^p |b^i_{jk}|.$$

Démonstration. Donnons la démonstration pour la norme ℓ_1. La $i$ème composante de $B(x,y)$ peut être majorée comme suit:

$$|B_i(x,y)| \leq \Big|\sum_{j=1}^n\sum_{k=1}^p b^i_{jk}x_jy_k\Big| \leq \max_{j,k}|b^i_{jk}|\Big(\sum_{j=1}^n|x_j|\Big)\Big(\sum_{k=1}^p|y_k|\Big).$$

On obtient donc

$$\|B(x,y)\|_1 = \sum_{i=1}^m |B_i(x,y)| \leq \sum_{i=1}^m \max_{j,k}|b^i_{jk}|\cdot\|x\|_1\cdot\|y\|_1$$

et on en déduit la majoration pour $\|B\|_1$.

La démonstration pour la norme maximum est similaire et celle pour la norme euclidienne utilise plusieurs fois l'inégalité de Cauchy-Schwarz. □

Applications multilinéaires Une application $M:\mathbb{R}^{n_1}\times\cdots\times\mathbb{R}^{n_k}\to\mathbb{R}^m$ s'appelle *multilinéaire* si elle est linéaire dans chacun des arguments. Les notations et résultats pour des applications bilinéaires se généralisent sans difficulté.

Une applications multilinéaire est toujours sous la forme

$$M(x^{[1]},\ldots,x^{[k]}) = \Big(\sum_{j_1=1}^{n_1}\cdots\sum_{j_k=1}^{n_k} m^i_{j_1,\ldots,j_k} x^{[1]}_{j_1}\cdot\ldots\cdot x^{[k]}_{j_k}\Big)_{i=1}^m$$

et sa norme est définie par

$$\|M\| = \sup_{\|x^{[1]}\|\leq 1,\ldots,\|x^{[k]}\|\leq 1}\|M(x^{[1]},\ldots,x^{[k]})\| = \sup_{x^{[1]}\neq 0,\ldots,x^{[k]}\neq 0}\frac{\|M(x^{[1]},\ldots,x^{[k]})\|}{\|x^{[1]}\|\cdot\ldots\cdot\|x^{[k]}\|}. \tag{6.5}$$

On a des majorations comme dans le théorème 6.2. Par exemple,

$$\|M\|_\infty = \max_{i=1,\ldots,m}\sum_{j_1=1}^{n_1}\cdots\sum_{j_k=1}^{n_k}|m^i_{j_1,\ldots,j_k}|.$$

Une des raisons pour laquelle on travaille avec des applications multilinéaires est justement d'éviter des expressions avec beaucoup de sommes et d'indices.

II.7 Exercices

Chapter III

Calcul différentiel (plusieurs variables)

Le but de ce chapitre est d'introduire la notion de différentiabilité pour des fonctions à plusieurs variables. Comme la division par le vecteur $x - x_0$ n'a pas de sens, il n'y a pas de manière évidente d'étendre la définition $f'(x_0) = \lim_{x\to x_0} \frac{f(x)-f(x_0)}{x-x_0}$.

Nous suivrons de près la présentation des paragraphes IV.3 et IV.4 du livre "L'analyse au fil de l'histoire" de Hairer & Wanner (Springer-Verlag 2000).

III.1 Dérivées partielles

Pour une fonction $f : U \to \mathbb{R}$, $U \subset \mathbb{R}^n$, nous fixons toutes les variables sauf une et nous considérons f comme une fonction de cette variable. Nous pouvons alors appliquer la définition bien connue pour une fonction à une variable. Considérons plus en détail le cas d'une fonction à deux variables.

Une fonction à deux variables Considérons par exemple une fonction $y = f(x_1, x_2)$ qui est définie dans un voisinage de (x_{10}, x_{20}). Alors, les limites

$$\begin{aligned} \lim_{h\to 0} \frac{f(x_{10}+h, x_{20}) - f(x_{10}, x_{20})}{h} &=: \frac{\partial f}{\partial x_1}(x_{10}, x_{20}) \\ \lim_{h\to 0} \frac{f(x_{10}, x_{20}+h) - f(x_{10}, x_{20})}{h} &=: \frac{\partial f}{\partial x_2}(x_{10}, x_{20}), \end{aligned} \tag{1.1}$$

s'appellent les *dérivées partielles* de f par rapport à x_1 et x_2, respectivement. Parmi d'autres notations on utilise $f_{x_i}(x_{10}, x_{20})$, $D_i f(x_{10}, x_{20})$, $\partial_i f(x_{10}, x_{20})$.

Géométriquement, les dérivées partielles peuvent être interprétées comme suit : la fonction $y = f(x_1, x_2)$ définit une surface dans $\mathbb{R}^3$ (avec des coordonnées x_1, x_2, et y), dont l'intersection avec le plan $x_2 = x_{20}$ est une courbe $x_1 \mapsto f(x_1, x_{20})$. La dérivée partielle $\partial f/\partial x_1$ est la pente de cette courbe, et

$$y = f(x_{10}, x_{20}) + \frac{\partial f}{\partial x_1}(x_{10}, x_{20})(x_1 - x_{10}) \tag{1.2}$$

est l'équation de la tangente à cette courbe en (x_{10}, x_{20}). De la même manière, la tangente à la courbe $x_2 \mapsto f(x_{10}, x_2)$ a pour équation $y = f(x_{10}, x_{20}) + \partial f/\partial x_2(x_{10}, x_{20})(x_2 - x_{20})$, et le plan déterminé par ces deux tangentes est donné par

$$y = f(x_{10}, x_{20}) + \frac{\partial f}{\partial x_1}(x_{10}, x_{20})(x_1 - x_{10}) + \frac{\partial f}{\partial x_2}(x_{10}, x_{20})(x_2 - x_{20}). \tag{1.3}$$

Une fonction $f(x_1, x_2)$ sera dite *différentiable* en (x_{10}, x_{20}) si le plan (1.3) est une “bonne” approximation de $f(x_1, x_2)$ dans un voisinage de (x_{10}, x_{20}), et pas seulement le long des lignes $x_1 = x_{10}$ et $x_2 = x_{20}$.

Exemple 1.1 La surface définie par $y = e^{-x_1^2 - x_2^2}$ est représentée sur la figure III.1. Nous voyons également les courbes $x_1 \mapsto f(x_1, x_{20})$ et $x_2 \mapsto f(x_{10}, x_2)$ passant par le point $(x_{10}, x_{20}) = (0.3, 0.1)$. En évaluant les dérivées partielles

$$\frac{\partial f}{\partial x_1}(x_1, x_2) = -2x_1 e^{-x_1^2 - x_2^2}, \qquad \frac{\partial f}{\partial x_2}(x_1, x_2) = -2x_2 e^{-x_1^2 - x_2^2}$$

en ce point, nous obtenons les tangentes en ces courbes au point (x_{10}, x_{20}) à l'aide de la formule (1.2) ainsi que le plan tangent à l'aide de (1.3).

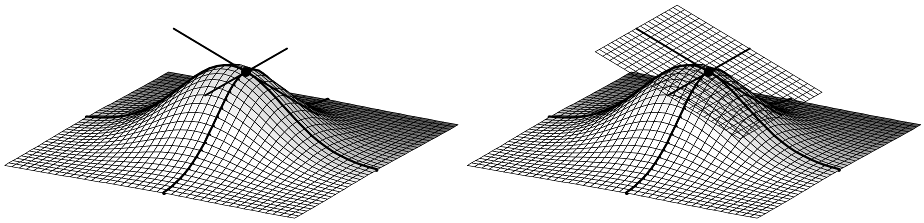


Figure III.1: Plan tangent à la surface $y = e^{-x_1^2 - x_2^2}$

Deux fonctions à deux variables Dans le cas d'une fonction $f : \mathbb{R}^2 \to \mathbb{R}^2$, c.-à-d.

$$y_1 = f_1(x_1, x_2), \qquad y_2 = f_2(x_1, x_2), \tag{1.4}$$

nous écrivons (1.3) pour chacune des deux fonctions :

$$\begin{aligned} y_1 &= f_1(x_{10}, x_{20}) + \frac{\partial f_1}{\partial x_1}(x_{10}, x_{20})(x_1 - x_{10}) + \frac{\partial f_1}{\partial x_2}(x_{10}, x_{20})(x_2 - x_{20}), \\ y_2 &= f_2(x_{10}, x_{20}) + \frac{\partial f_2}{\partial x_1}(x_{10}, x_{20})(x_1 - x_{10}) + \frac{\partial f_2}{\partial x_2}(x_{10}, x_{20})(x_2 - x_{20}). \end{aligned} \tag{1.5}$$

Il est commode d'écrire cette formule en notation vectorielle :

$$y = f(x_0) + f'(x_0)(x - x_0), \tag{1.6}$$

où $f'(x_0)$ est une matrice, appelée *matrice jacobienne* :

$$f'(x_0) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x_0) & \frac{\partial f_1}{\partial x_2}(x_0) \\ \frac{\partial f_2}{\partial x_1}(x_0) & \frac{\partial f_2}{\partial x_2}(x_0) \end{pmatrix}. \tag{1.7}$$

Cette notation nous permettra de transposer la plupart des formules d'une variable au cas de plusieurs variables.

Exemple 1.2 Considérons la fonction $f : \mathbb{R}^2 \to \mathbb{R}^2$ définie par

$$f(x) = \begin{pmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{pmatrix} = \begin{pmatrix} x_1 - 1.1\sin(x_1 + x_2 + 1) + 1.1\sin 1 \\ x_2 + 0.1x_1 - 0.8\cos(x_1 - x_2 + 1) + 0.8\cos 1 \end{pmatrix}. \tag{1.8}$$

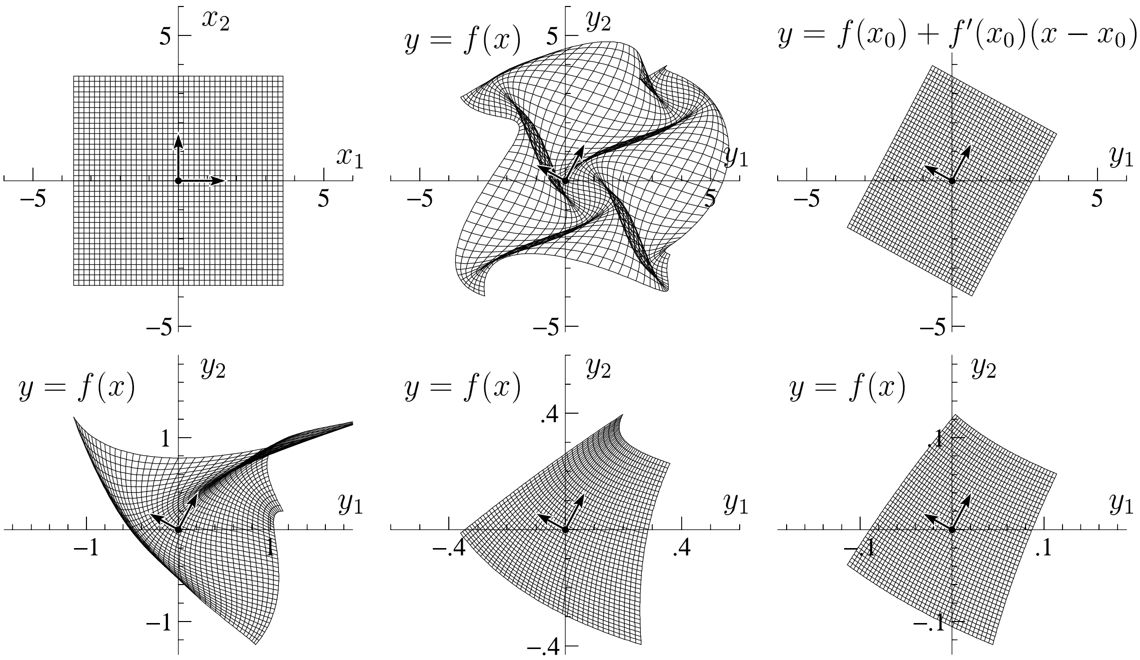


Figure III.2: Graphique de l'application (1.8)

Cette fonction envoie l'origine $(x_1, x_2) = (0, 0)$ sur le point $(y_1, y_2) = (0, 0)$, les droites sur des courbes, et les petits carrés sur des ensembles qui ressemblent à des parallélogrammes (voir la figure III.2, dessin en haut au milieu). La matrice jacobienne pour (1.8) est

$$f'(x) = \begin{pmatrix} 1 - 1.1\cos(x_1 + x_2 + 1) & -1.1\cos(x_1 + x_2 + 1) \\ 0.1 + 0.8\sin(x_1 - x_2 + 1) & 1 - 0.8\sin(x_1 - x_2 + 1) \end{pmatrix}, \tag{1.9}$$

et l'équation (1.5) devient pour $x_0 = (0, 0)^t$, $y_0 = f(x_0)$:

$$\begin{pmatrix} y_1 - y_{10} \\ y_2 - y_{20} \end{pmatrix} = \begin{pmatrix} 1 - 1.1\cos 1 & -1.1\cos 1 \\ 0.1 + 0.8\sin 1 & 1 - 0.8\sin 1 \end{pmatrix} \begin{pmatrix} x_1 - x_{10} \\ x_2 - x_{20} \end{pmatrix}. \tag{1.10}$$

Cette application linéaire est illustrée dans le dessin en haut à droite de la figure III.2. Les trois dessins de la deuxième ligne de la figure III.2 sont des agrandissements de cette application non linéaire proche de l'origine. Nous voyons qu'elle est, *dans un petit voisinage de* x_0, bien approchée par l'application linéaire définie par la matrice jacobienne.

III.2 Différentiabilité

Considérons une fonction

$$f : U \to \mathbb{R}^m, \qquad U \subset \mathbb{R}^n \tag{2.1}$$

et supposons que $x_0 \in U$ soit un *point intérieur* de U (c.-à-d. U est un voisinage de x_0).

Définition 2.1 (Stolz 1887, Fréchet 1906) *La fonction (2.1) est différentiable en x_0 s'il existe une application linéaire $f'(x_0) : \mathbb{R}^n \to \mathbb{R}^m$ et une fonction $r : U \to \mathbb{R}^m$, continue en x_0 et satisfaisant $r(x_0) = 0$, telles que*

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + r(x)\|x - x_0\|. \tag{2.2}$$

Remarque. Si une fonction est différentiable en x_0, elle est continue en ce point. De plus, toutes ses dérivées partielles existent en x_0. C'est une conséquence du fait que, pour $x - x_0 = he_j$ (où $e_j = (0, \ldots, 0, 1, 0, \ldots, 0)^t$ avec la $j$ème composante égale à 1), l'équation (2.2) devient

$$\frac{f(x_0 + he_j) - f(x_0)}{h} = f'(x_0)e_j + r(x_0 + he_j)\frac{|h|}{h}. \tag{2.3}$$

Comme $r(x)$ est continue en x_0, la limite de cette expression pour $h \to 0$ existe et est égale à

$$\frac{\partial f}{\partial x_j}(x_0) = f'(x_0)e_j, \qquad \text{donc} \qquad f'(x_0) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x_0) & \cdots & \frac{\partial f_1}{\partial x_n}(x_0) \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(x_0) & \cdots & \frac{\partial f_m}{\partial x_n}(x_0) \end{pmatrix} \tag{2.4}$$

(ici, $f(x) = (f_1(x), \ldots, f_m(x))^t$). Par conséquent, l'application linéaire de la définition 2.1 est unique et donnée par la matrice jacobienne (2.4).

Lemme 2.2 (caractérisation de Carathéodory) *La fonction $f(x)$ de (2.1) est différentiable en x_0 si et seulement s'il existe une fonction $\varphi(x)$ à valeurs matricielles, dépendant de x_0 et continue en x_0, telle que*

$$f(x) = f(x_0) + \varphi(x)(x - x_0). \tag{2.5}$$

La dérivée de $f(x)$ en x_0 est donnée par $f'(x_0) = \varphi(x_0)$.

Démonstration. Pour une fonction donnée $\varphi(x)$, nous posons

$$f'(x_0) := \varphi(x_0), \qquad r(x) := (\varphi(x) - \varphi(x_0))\frac{(x - x_0)}{\|x - x_0\|},$$

et nous voyons que (2.2) est vérifiée. Comme $(x - x_0)/\|x - x_0\|$ est borné par 1, il suit de la continuité de $\varphi(x)$ en x_0 que $r(x) \to 0$ pour $x \to x_0$.

D'autre part, supposons que (2.2) soit vrai. Nous définissons $\varphi(x_0) := f'(x_0)$, et pour $x \neq x_0$,

$$\varphi(x) := f'(x_0) + r(x)\frac{(x - x_0)^t}{\|x - x_0\|} \tag{2.6}$$

(observons que le produit du vecteur colonne $r(x)$ avec le vecteur ligne $(x - x_0)^t$ est une matrice), et nous obtenons $\varphi(x)(x - x_0) = f'(x_0)(x - x_0) + r(x)\|x - x_0\|$. La fonction $\varphi(x)$ est continue en x_0 car, par la définition de la norme matricielle et en utilisant l'inégalité de Cauchy-Schwarz, $\|\varphi(x) - f'(x_0)\| \leq \|r(x)\|$ et $\|r(x)\| \to 0$ pour $x \to x_0$. □

Le résultat suivant donne une condition suffisante pour la différentiabilité qui peut être vérifiée en considérant seulement des dérivées partielles.

Théorème 2.3 *Considérons une fonction $f : U \to \mathbb{R}$ et un $x_0 \in U$ (point intérieur). Si toutes les dérivées partielles $\partial f/\partial x_i$ existent dans un voisinage de x_0 et sont continues en x_0, alors f est différentiable en x_0.*

Démonstration. Nous expliciterons la démonstration pour le cas $n = 2$. La généralisation à un n arbitraire est immédiate. L'idée est d'écrire $f(x) - f(x_0)$ comme

$$f(x_1, x_2) - f(x_{10}, x_{20}) = (f(x_1, x_2) - f(x_{10}, x_2)) + (f(x_{10}, x_2) - f(x_{10}, x_{20}))$$

et d'appliquer le théorème de Lagrange à chacune des différences. Ceci donne

$$f(x_1,x_2)-f(x_{10},x_{20})=\frac{\partial f}{\partial x_1}(\xi_1,x_2)(x_1-x_{10})+\frac{\partial f}{\partial x_2}(x_{10},\xi_2)(x_2-x_{20})$$

avec ξ_1 entre x_{10} et x_1, et ξ_2 entre x_{20} et x_2. En posant

$$\varphi(x_1,x_2)=\Big(\frac{\partial f}{\partial x_1}(\xi_1,x_2),\frac{\partial f}{\partial x_2}(x_{10},\xi_2)\Big),$$

nous avons établi (2.5). La continuité de $\varphi(x)$ en x_0 est une conséquence des hypothèses. □

Par la définition 2.1, une fonction $f(x)=(f_1(x),\ldots,f_m(x))^t$ est différentiable en x_0 si et seulement si $f_i(x)$ est différentiable en x_0 pour tout $i=1,\ldots,m$. Une conséquence du théorème 2.3 est que les fonctions dont les composantes sont des polynômes en $x_1,\ldots,x_n$, des fonctions rationnelles ou des fonctions élémentaires sont différentiables aux points où elles sont bien définies.

Exemple 2.4 (Une fonction discontinue dont les dérivées partielles existent partout) Considérons la fonction $f:\mathbb{R}^2\to\mathbb{R}$ donnée par

$$f(x_1,x_2)=\begin{cases}\dfrac{x_1x_2}{x_1^2+x_2^2} & \text{si } x_1^2+x_2^2>0\\ 0 & \text{si } x_1=x_2=0\end{cases} \tag{2.7}$$

(voir la figure I.9). Ses dérivées partielles sont nulles à l'origine car $f(x_1,0)=0$ pour tout x_1 et $f(0,x_2)=0$ pour tout x_2. En dehors de l'origine, l'existence des dérivées partielles est évidente. Néanmoins, la fonction (2.7) n'est pas continue à l'origine (voir l'exemple I.4.3).

La dérivée en chaîne Considérons deux fonctions

$$\begin{array}{ccccc}\mathbb{R}^n & \xrightarrow{f} & \mathbb{R}^m & \xrightarrow{g} & \mathbb{R}^p\\ x & \longmapsto & y & \longmapsto & z\end{array}$$

et étudions la différentiabilité de la fonction composée $(g\circ f)(x)=g(f(x))$. Nous utilisons la caractérisation de Carathéodory (lemme 2.2). En supposant que f soit différentiable en x_0 et que g soit différentiable en $y_0=f(x_0)$, nous avons

$$f(x)=f(x_0)+\varphi(x)(x-x_0),\qquad g(y)=g(y_0)+\psi(y)(y-y_0).$$

En posant $y=f(x)$, $y_0=f(x_0)$ et en introduisant la première équation dans la deuxième, nous obtenons

$$g(f(x))=g(f(x_0))+\psi(f(x))\varphi(x)(x-x_0). \tag{2.8}$$

Puisque le produit $\psi(f(x))\varphi(x)$ est continu en x_0, la dérivée de $g\circ f$ est égale à ce produit évalué en x_0, i.e.

$$(g\circ f)'(x_0)=g'(y_0)\cdot f'(x_0). \tag{2.9}$$

En coordonnées, le produit (2.9) s'écrit

$$\frac{\partial z_i}{\partial x_k}=\sum_{j=1}^m\frac{\partial z_i}{\partial y_j}\cdot\frac{\partial y_j}{\partial x_k}\,. \tag{2.10}$$

Exemple 2.5 Supposons que le mouvement d'un corps soit décrit en coordonnées polaires par $f(t) = (r(t), \varphi(t))^t$. Si nous voulons connaître la vitesse en coordonnées cartésiennes

$$\begin{pmatrix} x \\ y \end{pmatrix} = g(r, \varphi) = \begin{pmatrix} r\cos\varphi \\ r\sin\varphi \end{pmatrix}, \tag{2.11}$$

il nous faut dériver x et y par rapport à t. Comme la matrice jacobienne de (2.11) est donnée par

$$g'(r, \varphi) = \begin{pmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{pmatrix}, \tag{2.12}$$

il suit de (2.9) que

$$\dot{x} = \cos\varphi \cdot \dot{r} - r\sin\varphi \cdot \dot{\varphi}, \qquad \dot{y} = \sin\varphi \cdot \dot{r} + r\cos\varphi \cdot \dot{\varphi}$$

(la dérivée par rapport au temps t est désignée par un point). Ceci nous permet, par exemple, de calculer l'énergie cinétique

$$T(t) = \frac{m}{2}(\dot{x}^2 + \dot{y}^2) = \frac{m}{2}(\dot{r}^2 + r^2\dot{\varphi}^2).$$

III.3 Dérivées d'ordre supérieur

Considérons tout d'abord une fonction $f(x, y)$ de deux variables. Ses dérivées partielles, par exemple $\partial f/\partial x$, sont aussi des fonctions de deux variables et nous pouvons calculer itérativement leurs dérivées partielles comme dans le diagramme suivant :

$$\begin{array}{ccccccc}
f(x,y) & \xrightarrow{\frac{\partial}{\partial x}} & & \dfrac{\partial f}{\partial x} & \xrightarrow{\frac{\partial}{\partial x}} & \dfrac{\partial^2 f}{\partial x^2} & \cdots \\
\frac{\partial}{\partial y}\Big\downarrow & & & \frac{\partial}{\partial y}\Big\downarrow & & & \\
\dfrac{\partial f}{\partial y} & \xrightarrow{\frac{\partial}{\partial x}} & \dfrac{\partial^2 f}{\partial x \partial y} \overset{?}{=} & \dfrac{\partial^2 f}{\partial y \partial x} & \xrightarrow{\frac{\partial}{\partial x}} & \cdots & \\
\frac{\partial}{\partial y}\Big\downarrow & & & \frac{\partial}{\partial y}\Big\downarrow & & & \\
\dfrac{\partial^2 f}{\partial y^2} & \xrightarrow{\frac{\partial}{\partial x}} & \dfrac{\partial^3 f}{\partial x \partial y^2} \overset{?}{=} & \dfrac{\partial^3 f}{\partial y^2 \partial x} & \xrightarrow{\frac{\partial}{\partial x}} & \cdots\ . &
\end{array}$$

La question est de savoir si ces dérivées dépendent de l'ordre de différentiation.

Exemple 3.1 Considérons la fonction $f(x, y) = \sqrt{x^2 + 5y^2}$ et calculons ses dérivées partielles (pour $x^2 + 5y^2 > 0$) :

$$\begin{aligned}
\frac{\partial f}{\partial x}(x,y) &= \frac{x}{\sqrt{x^2+5y^2}}, & \frac{\partial^2 f}{\partial y \partial x}(x,y) &= \frac{-5xy}{(x^2+5y^2)^{3/2}}, \\
\frac{\partial f}{\partial y}(x,y) &= \frac{5y}{\sqrt{x^2+5y^2}}, & \frac{\partial^2 f}{\partial x \partial y}(x,y) &= \frac{-5xy}{(x^2+5y^2)^{3/2}}.
\end{aligned}$$

Nous obtenons que

$$\frac{\partial^2 f}{\partial y \partial x}(x,y) = \frac{\partial^2 f}{\partial x \partial y}(x,y) \tag{3.1}$$

pour cette fonction. Un calcul avec d'autres fonctions (par exemple avec des polynômes) laisse deviner que (3.1) est toujours vrai. Cependant, (3.1) n'est pas vrai sans hypothèse supplémentaire, comme on peut le voir à l'aide du contre-exemple suivant.

Exemple 3.2 (Contre-exemple) Pour trouver une fonction qui ne vérifie pas (3.1), nous considérons une fonction

$$f(x,y) = x\,y\,g(x,y), \tag{3.2}$$

où $g(x,y)$ est bornée (mais pas nécessairement continue) dans un voisinage de l'origine. Pour cette fonction, on a

$$\frac{\partial f}{\partial x}(0,y) = \lim_{x\to 0}\frac{f(x,y)-f(0,y)}{x} = \lim_{x\to 0} y\,g(x,y).$$

La dérivée de cette expression par rapport à y est

$$\frac{\partial^2 f}{\partial y\partial x}(0,0) = \lim_{y\to 0}\Big(\lim_{x\to 0} g(x,y)\Big), \tag{3.3}$$

à condition que cette limite existe. De même, on a

$$\frac{\partial^2 f}{\partial x\partial y}(0,0) = \lim_{x\to 0}\Big(\lim_{y\to 0} g(x,y)\Big). \tag{3.4}$$

Il suffit de choisir une fonction $g(x,y)$ pour laquelle les limites dans (3.3) et (3.4) sont différentes. C'est le cas de

$$g(x,y) = \frac{x^2-y^2}{x^2+y^2} \qquad \text{si} \quad x^2+y^2>0, \tag{3.5}$$

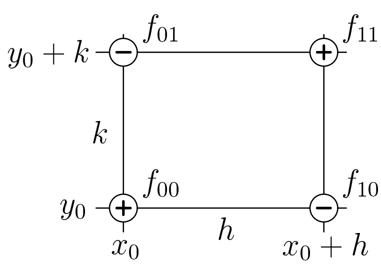
pour laquelle $\lim_{x\to 0} g(x,y) = -1$ pour tout $y \neq 0$ et $\lim_{y\to 0} g(x,y) = +1$ pour tout $x \neq 0$. Par conséquent, les dérivées partielles

$$\frac{\partial^2 f}{\partial y\partial x}(0,0) = -1 \qquad \text{et} \qquad \frac{\partial^2 f}{\partial x\partial y}(0,0) = 1$$

sont différentes pour la fonction définie par (3.2) et (3.5).

Théorème 3.3 *Considérons une fonction* $f : \mathbb{R}^2 \to \mathbb{R}$ *dont les dérivées partielles* $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$, $\frac{\partial^2 f}{\partial y\partial x}$ *existent dans un voisinage de* (x_0,y_0), *avec* $\frac{\partial^2 f}{\partial y\partial x}$ *continue en* (x_0,y_0). *Alors,* $\frac{\partial^2 f}{\partial x\partial y}$ *existe en* (x_0,y_0) *et*

$$\frac{\partial^2 f}{\partial x\partial y}(x_0,y_0) = \frac{\partial^2 f}{\partial y\partial x}(x_0,y_0).$$



Démonstration. L'idée est de considérer un petit rectangle de côtés h et k. On désigne les valeurs de f aux sommets de ce rectangle par f_{00}, f_{01}, f_{10} et f_{11}. Les dérivées partielles sont approximativement données par

$$\frac{\partial f}{\partial x}(x_0,y_0) \approx \frac{f_{10}-f_{00}}{h}, \qquad \frac{\partial f}{\partial x}(x_0,y_0+k) \approx \frac{f_{11}-f_{01}}{h}.$$

On en déduit que

$$\frac{\partial^2 f}{\partial y\partial x} \approx \frac{\frac{\partial f}{\partial x}(x_0,y_0+k) - \frac{\partial f}{\partial x}(x_0,y_0)}{k} \approx \frac{f_{11}-f_{01}-f_{10}+f_{00}}{h\cdot k}, \tag{3.6}$$

et de même

$$\frac{\partial^2 f}{\partial x\partial y} \approx \frac{\frac{\partial f}{\partial y}(x_0+h,y_0) - \frac{\partial f}{\partial y}(x_0,y_0)}{h} \approx \frac{f_{11}-f_{10}-f_{01}+f_{00}}{k\cdot h}. \tag{3.7}$$

Les membres de droite dans (3.6) et (3.7) sont identiques et l'affirmation du théorème est plausible.

Pour une démonstration rigoureuse de ce théorème nous vous référons à la page 318 du livre "Analyse au fil de l'histoire". □

L'application répétée de ce théorème permet d'échanger des dérivées d'ordre supérieur. Par exemple,

$$\frac{\partial}{\partial x}\underbrace{\frac{\partial}{\partial y}\frac{\partial}{\partial x}}_{\nwarrow\nearrow}\underbrace{\frac{\partial}{\partial x}\frac{\partial}{\partial y}f}_{g} = \frac{\partial}{\partial x}\frac{\partial}{\partial x}\underbrace{\frac{\partial}{\partial y}\frac{\partial}{\partial x}}_{\nwarrow\nearrow}\underbrace{\frac{\partial}{\partial y}f}_{g} = \frac{\partial}{\partial x}\frac{\partial}{\partial x}\frac{\partial}{\partial x}\frac{\partial}{\partial y}\frac{\partial}{\partial y}f = \ldots\ .$$

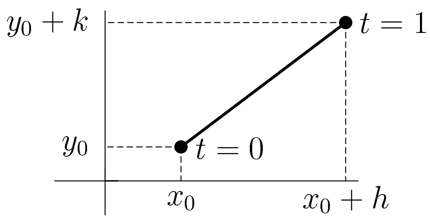
Cette manière de procéder est aussi valable pour des fonctions de plus de deux variables. En effet, nous échangeons toujours *deux dérivées* partielles à la fois, en maintenant constantes les autres variables.

III.4 Série de Taylor

Rappelons que la série de Taylor d'une fonction $g(t)$ à une variable est donnée par

$$g(t) \;=\; g(0) + g'(0)\,t + g''(0)\,\frac{t^2}{2!} + +g'''(0)\,\frac{t^3}{3!} + \ldots\ . \tag{4.1}$$

Il s'agit de généraliser cette formule à des fonctions de plusieurs variables. Pour éviter une notation trop lourde, considérons d'abord le cas de deux variables.



Deux variables L'idée est de ramener le problème à une variable en reliant les points (x_0, y_0) et $(x_0 + h, y_0 + k)$ par une droite. On considère alors la fonction

$$g(t) := f(x_0 + th, y_0 + tk)$$

et on applique la formule (4.1) avec $t = 1$. Il reste à calculer les dérivées de $g(t)$. Si $f(x, y)$ est suffisamment différentiable, la dérivée en chaîne nous donne

$$g'(t) = \frac{\partial f}{\partial x}(x_0 + th, y_0 + tk)\,h + \frac{\partial f}{\partial y}(x_0 + th, y_0 + tk)\,k \tag{4.2}$$

et une différentiation de plus donne

$$g''(t) = \frac{\partial^2 f}{\partial x^2}(\cdot)hh + \frac{\partial^2 f}{\partial y \partial x}(\cdot)hk + \frac{\partial^2 f}{\partial x \partial y}(\cdot)kh + \frac{\partial^2 f}{\partial y^2}(\cdot)kk, \tag{4.3}$$

où l'argument des dérivées partielles de f, que nous avons omis, est $(x_0 + th, y_0 + tk)$. Les deux termes centraux dans (4.3) sont égaux par le théorème 3.3 (en différentiant encore, on ferait apparaître des coefficients binomiaux). En insérant ces dérivées de $g(t)$ dans (4.1), il vient

$$\begin{aligned} f(x_0 + h, y_0 + k) \;&=\; f(x_0, y_0) + \frac{\partial f}{\partial x}(x_0, y_0)h + \frac{\partial f}{\partial y}(x_0, y_0)k \\ &+ \frac{1}{2}\Big(\frac{\partial^2 f}{\partial x^2}(x_0, y_0)h^2 + 2\frac{\partial^2 f}{\partial x \partial y}(x_0, y_0)hk + \frac{\partial^2 f}{\partial y^2}(x_0, y_0)k^2\Big) \\ &+ \frac{1}{6}\Big(\frac{\partial^3 f}{\partial x^3}(x_0, y_0)h^3 + 3\frac{\partial^3 f}{\partial x^2 \partial y}(x_0, y_0)h^2k + 3\frac{\partial^3 f}{\partial x \partial y^2}(x_0, y_0)hk^2 + \frac{\partial^3 f}{\partial y^3}(x_0, y_0)k^3\Big) + \ldots \end{aligned} \tag{4.4}$$

Exemple 4.1 Considérons la fonction $f(x,y) = e^{-x^2-y^2}$ (voir aussi l'exemple 1.1), dont les dérivées partielles sont données par

$$\frac{\partial f}{\partial x}(x,y) = -2xe^{-x^2-y^2}, \qquad \frac{\partial f}{\partial y}(x,y) = -2ye^{-x^2-y^2},$$
$$\frac{\partial^2 f}{\partial x^2}(x,y) = (4x^2-2)e^{-x^2-y^2}, \qquad \frac{\partial^2 f}{\partial y^2}(x,y) = (4y^2-2)e^{-x^2-y^2},$$
$$\frac{\partial^2 f}{\partial x \partial y}(x,y) = \frac{\partial^2 f}{\partial y \partial x}(x,y) = 4xye^{-x^2-y^2}.$$

En négligeant les termes d'ordres trois et supérieurs dans la formule (4.4) et en posant $x_0 = 0.3$, $y_0 = 0.1$, nous obtenons l'approximation quadratique

$$f(0.3+h, 0.1+k) \approx e^{-0.1}(1 - 0.6h - 0.2k - 1.64h^2 + 0.12hk - 1.96k^2)$$

qui représente un paraboloïde. La figure III.3 compare cette approximation avec la fonction $f(x,y)$ et avec le plan tangent.

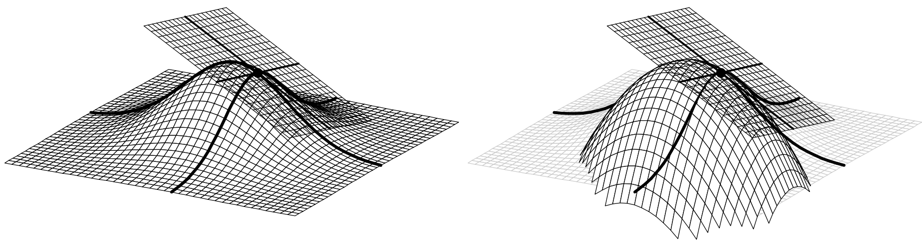


Figure III.3: Approximation de Taylor d'ordre un, comparée à l'ordre deux pour $f(x,y) = e^{-x^2-y^2}$

Série de Taylor pour n variables Nous généralisons maintenant ces formules aux fonctions

$$f : \mathbb{R}^n \to \mathbb{R}^m,$$

où $f(x) = (f_1(x), \ldots, f_m(x))^t$ est composée de m fonctions réelles de $x \in \mathbb{R}^n$. Nous fixons $x_0 \in \mathbb{R}^n$, $h \in \mathbb{R}^n$ et nous appliquons la formule (4.1) à la fonction $g(t) := f_i(x_0 + th)$. Ceci donne

$$\begin{aligned} f_i(x_0+h) \;=\; f_i(x_0) \;&+\; \sum_{j=1}^n \frac{\partial f_i}{\partial x_j}(x_0)h_j + \frac{1}{2!}\sum_{j=1}^n \sum_{k=1}^n \frac{\partial^2 f_i}{\partial x_j \partial x_k}(x_0)\, h_j h_k \\ &+\; \frac{1}{3!}\sum_{j=1}^n \sum_{k=1}^n \sum_{\ell=1}^n \frac{\partial^3 f_i}{\partial x_j \partial x_k \partial x_\ell}(x_0)\, h_j h_k h_\ell \;+\; \ldots\;. \end{aligned} \tag{4.5}$$

On peut aussi aller plus loin et écrire (formellement, sans considérer la convergence)

$$f_i(x_0+h) = f_i(x_0) + \sum_{q=1}^\infty \frac{1}{q!} \sum_{j_1=1}^n \sum_{j_2=1}^n \ldots \sum_{j_q=1}^n \frac{\partial^q f_i(x_0)}{\partial x_{j_1} \partial x_{j_2} \ldots \partial x_{j_q}} h_{j_1} \ldots h_{j_q}.$$

Ces formules, assez encombrantes, nécessitent une notation plus compacte.

Le terme linéaire dans (4.5) n'est autre que le $i$ème élément du produit $f'(x_0)h$ (de la matrice jacobienne avec le vecteur h). Pour simplifier le terme quadratique, nous considérons l'*application bilinéaire* $f''(x) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m$ appliquée à un couple de vecteurs u et v, dont la $i$ème composante est définie par

$$\Big(f''(x)(u,v)\Big)_i := \sum_{j=1}^n \sum_{k=1}^n \frac{\partial^2 f_i}{\partial x_j \partial x_k}(x)\, u_j v_k. \tag{4.6}$$

Le terme quadratique dans (4.5) est donc le $i$ème élément du vecteur $f''(x_0)(h,h)$. On peut continuer de cette manière à interpréter les dérivées supérieures comme *applications multilinéaires*. Par exemple, $f'''(x) : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m$ est définie par

$$\Big(f'''(x)(u,v,w)\Big)_i := \sum_{j=1}^n \sum_{k=1}^n \sum_{\ell=1}^n \frac{\partial^3 f_i}{\partial x_j \partial x_k \partial x_\ell}(x)\, u_j v_k w_\ell. \tag{4.7}$$

Avec cette notation, la formule (4.5) devient

$$f(x_0+h) = f(x_0) + f'(x_0)h + \frac{1}{2!} f''(x_0)(h,h) + \frac{1}{3!} f'''(x_0)(h,h,h) + \ldots\ . \tag{4.8}$$

Formule de Taylor avec reste Même si toutes les dérivées de $g(t)$ en $t = 0$ existent, la convergence de la série (4.1) n'est pas garantie. Il est alors souvent utile de considérer une série tronquée et d'estimer le reste. Rappelons, par exemple, que

$$g(t) \ = \ g(0) + g'(0)\,t + g''(0)\,\frac{t^2}{2!} + R_3 \qquad \text{avec} \qquad R_3 = \int_0^t \frac{(t-s)^2}{2!}\, g'''(s)\,ds\,. \tag{4.9}$$

Pour une fonction $f : \mathbb{R}^n \to \mathbb{R}^m$ cela devient

$$f(x_0+h) = f(x_0) + f'(x_0)h + \frac{1}{2!} f''(x_0)(h,h) + R_3, \quad R_3 = \int_0^1 \frac{(1-t)^2}{2!}\, f'''(x_0+th)(h,h,h)\,dt. \tag{4.10}$$

Remarque. Pour une fonction vectorielle $g(t) = (g_1(t), \ldots, g_m(t))^t$, nous écrivons

$$\int_0^1 g(t)\,dt := \Big(\int_0^1 g_1(t)\,dt, \ldots, \int_0^1 g_m(t)\,dt\Big)^t. \tag{4.11}$$

Par la suite, nous utiliserons l'estimation

$$\Big\| \int_0^1 g(t)\,dt \Big\| \le \int_0^1 \|g(t)\|\,dt, \tag{4.12}$$

que l'on obtient en appliquant l'inégalité du triangle aux sommes de Riemann : $\| \sum_i g(\xi_i)\delta_i\| \le \sum_i \|g(\xi_i)\|\delta_i$.

Lemme 4.2 (estimation du reste) *Soit $f : \mathbb{R}^n \to \mathbb{R}^m$ trois fois continûment différentiable (c.-à-d. toutes les troisièmes dérivées partielles existent et sont continues). Le reste R_3 dans l'équation (4.10) vérifie alors*

$$\|R_3\| \le \frac{\|h\|^3}{3!} \sup_{t\in[0,1]} \|f'''(x_0+th)\|.$$

Démonstration. Les estimations (4.12) et $\|f'''(x)(u,v,w)\| \le \|f'''(x)\| \cdot \|u\| \cdot \|v\| \cdot \|w\|$ pour une application multilinéaire donnent

$$\|R_3\| \le \int_0^1 \frac{(1-t)^2}{2!}\, \|f'''(x_0+th)(h,h,h)\|\,dt \le \int_0^1 \frac{(1-t)^2}{2!} \sup_{t\in[0,1]} \|f'''(x_0+th)\| \cdot \|h\|^3\,dt$$

ce qui démontre l'affirmation. □

La norme $\|f'''(x_0+th)\|$ de l'application multilinéaire $f'''(x)$ peut être majorée par le formules du paragraphe II.6.

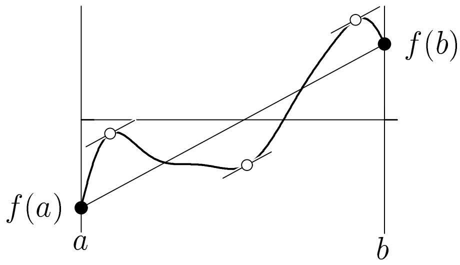
III.5 Théorème des accroissements finis

> La terminologie des ¡¡accroissements finis¿¿ s'explique par des raisons historiques: la notion d'accroissements ¡¡finis¿¿ s'oppose à celle d'accroissements ¡¡infinitésimaux¿¿ . . . (H. Cartan 1967)

Pour une fonction $f : [a, b] \to \mathbb{R}$, le théorème des accroissements finis (ou théorème de Lagrange) affirme qu'il existe un $\xi \in (a, b)$ tel que

$$f(b) - f(a) = f'(\xi)(b - a),$$

si f est continue sur $[a, b]$ et différentiable sur (a, b). Le but est de généraliser cette formule à plusieurs variables.



Un contre-exemple Pour une fonction $f : [a, b] \to \mathbb{R}^m$, l'affirmation sous cette forme n'est plus vraie. Un contre-exemple est la fonction $f(x) = (f_1(x), f_2(x))^t$ où $f_1(x) = \cos x$, $f_2(x) = \sin x$, et $[a, b] = [0, 2\pi]$. On a que $f(a) = f(b)$, mais il n'existe pas de ξ dans (a, b) avec $f'(\xi) = 0$. Par contre, on voit que l'inégalité $\|f(b) - f(a)\| \le \|f'(\xi)\| \cdot \|b - a\|$ reste vraie dans cet exemple.

Le cas $\mathrm{m} = 1$ Considérons une fonction $f : \mathbb{R}^n \to \mathbb{R}$ et soient $a \in \mathbb{R}^n$ et $b \in \mathbb{R}^n$ deux points donnés. L'idée est de relier ces points par une droite $x = a + t(b - a)$ (voir le début du paragraphe III.4) et de poser

$$g(t) := f(a + t(b - a)).$$

Si $f(x)$ est différentiable en chaque point du segment $\{a + t(b - a)\ ;\ t \in (0, 1)\}$, la fonction $g(t)$ est aussi différentiable et il suit de (2.9) que

$$g'(t) = f'(a + t(b - a))(b - a).$$

Comme $g(0) = f(a)$ et $g(1) = f(b)$, le théorème de Lagrange appliqué à la fonction $g(t)$ donne $g(1) - g(0) = g'(\tau)(1 - 0)$ et donc aussi

$$f(b) - f(a) = f'(\xi)(b - a), \tag{5.1}$$

où ξ est un point du segment reliant a et b. L'équation (5.1) ressemble à la formule du début de ce paragraphe, mais ici $f'(\xi)(b - a)$ est le produit scalaire de deux vecteurs.

Le cas général Soit maintenant $f : \mathbb{R}^n \to \mathbb{R}^m$. On peut appliquer (5.1) à chaque composante de $f(x)$ pour obtenir

$$\begin{pmatrix} f_1(b) - f_1(a) \\ \vdots \\ f_m(b) - f_m(a) \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\xi_1) & \dots & \frac{\partial f_1}{\partial x_n}(\xi_1) \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(\xi_m) & \dots & \frac{\partial f_m}{\partial x_n}(\xi_m) \end{pmatrix} \begin{pmatrix} b_1 - a_1 \\ \vdots \\ b_n - a_n \end{pmatrix}, \tag{5.2}$$

où tous les $\xi_i \in \mathbb{R}^n$ sont sur le segment reliant a et b. L'inconvénient de cette formule est que l'argument ξ_i est différent dans chaque ligne.

Théorème 5.1 (théorème des accroissements finis) *Soit $f : U \to \mathbb{R}^m$, $U \subset \mathbb{R}^n$ continue et différentiable en chaque point du segment "ouvert" $(a, b) := \{a + t(b - a)\ ;\ 0 < t < 1\}$ (on suppose que ces points sont des points intérieurs de U). Alors, on a*

$$\|f(b) - f(a)\| \le \sup_{\xi \in (a,b)} \|f'(\xi)\| \cdot \|b - a\|, \tag{5.3}$$

où $\|f'(\xi)\|$ est la norme matricielle de la matrice jacobienne.

Démonstration. L'idée est de considérer la fonction

$$g(t) := \sum_{i=1}^{m} c_i\, f_i(a + t(b-a)) = c^t f(a + t(b-a)), \tag{5.4}$$

où les coefficients $c_1, \ldots, c_m$ sont pour le moment arbitraires. La dérivée de $g(t)$ est

$$g'(t) = \sum_{i=1}^{m} c_i \sum_{j=1}^{n} \frac{\partial f_i}{\partial x_j}(a + t(b-a))(b_j - a_j) = c^t f'(a + t(b-a))(b-a).$$

Une application du théorème de Lagrange donne

$$c^t(f(b) - f(a)) = g(1) - g(0) = g'(\tau) = c^t f'(\xi)(b-a), \tag{5.5}$$

où $\xi = a + \tau(b-a)$ est sur le segment (a, b). Le choix astucieux $c = f(b) - f(a)$ rend maximal l'expression de gauche dans (5.5). En appliquant l'inégalité de Cauchy-Schwarz à l'expression de droite de l'équation (5.5), nous obtenons avec (II.5.3) :

$$\|f(b) - f(a)\|^2 \leq \|f(b) - f(a)\| \cdot \|f'(\xi)\| \cdot \|b - a\|.$$

En divisant par $\|f(b)-f(a)\|$, ceci donne (5.3) (observons que pour $\|f(b)-f(a)\| = 0$ l'affirmation (5.3) est évidente). □

III.6 Deux théorèmes importants de l'analyse

Le premier étudie l'existence et l'unicité de systèmes d'équations non linéaires. Le deuxième traite la résolution de fonctions implicites.

Théorème d'inversion locale. Rappelons que, pour une fonction $f : \mathbb{R} \to \mathbb{R}$ (continûment différentiable), la condition $f'(a) \neq 0$ implique que la fonction est monotone et donc bijective dans un voisinage de a. On cherche à généraliser ce résultat.

Pour une fonction $f : \mathbb{R}^n \to \mathbb{R}^n$ (continûment différentiable, c.-à-d. $f(x)$ est partout différentiable et sa dérivée $f'(x)$ est une fonction continue de x), considérons l'équation $f(x) = y$:

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_n) &= y_1 \\ f_2(x_1, x_2, \ldots, x_n) &= y_2 \\ &\cdots \\ f_n(x_1, x_2, \ldots, x_n) &= y_n\,. \end{aligned} \tag{6.1}$$

Elle représente un système de n équations à n inconnues. On aimerait savoir si, pour un $y \in V \subset \mathbb{R}^n$ donné, ce système possède une solution x. La solution, si elle existe, dépend-elle continûment (de manière différentiable) de y? On veut donc savoir si la fonction f possède (localement) un inverse.

Définition 6.1 Une fonction $f : \mathbb{R}^n \to \mathbb{R}^n$ est un *difféomorphisme local près de* a, s'il existe un voisinage U de a et un voisinage V de $b = f(a)$ tels que $f : U \to V$ soit bijective avec $f|_U$ et $f^{-1}|_V$ continûment différentiables.

Le théorème suivant donne un critère facile à vérifier qui implique qu'une fonction est un difféomorphisme près d'un point. Le résultat est donné sans démonstration.

Théorème 6.2 (théorème d'inversion locale) *Soit $f : \mathbb{R}^n \to \mathbb{R}^n$ continûment différentiable. Alors f est un difféomorphisme local près de $a \in \mathbb{R}^n$ si et seulement si la matrice Jacobienne*

$$f'(a) = \Big(\frac{\partial f_i}{\partial x_j}(a)\Big)_{i,j=1}^{n} \qquad \textit{est inversible.}$$

Exemple 6.3 Considérons la fonction $f : \mathbb{R}^2 \to \mathbb{R}^2$ définie par

$$y_1 = x_1 + 0.2\, x_2^3 \,, \qquad y_2 = 0.2\,(x_1 + 0.5)\, x_2^4 - 0.3\,(x_1 - 1)\, x_2 - 0.1\, x_1^3 \,. \tag{6.2}$$

Cette fonction est suffisamment compliquée pour qu'une résolution de l'équation $f(x) = y$ par rapport à $x = (x_1, x_2)^t$ soit impossible analytiquement. La fonction est illustrée dans la figure III.4, où on voit une partie d'un grillage dans le plan (x_1, x_2) ainsi que son image dans le plan (y_1, y_2).

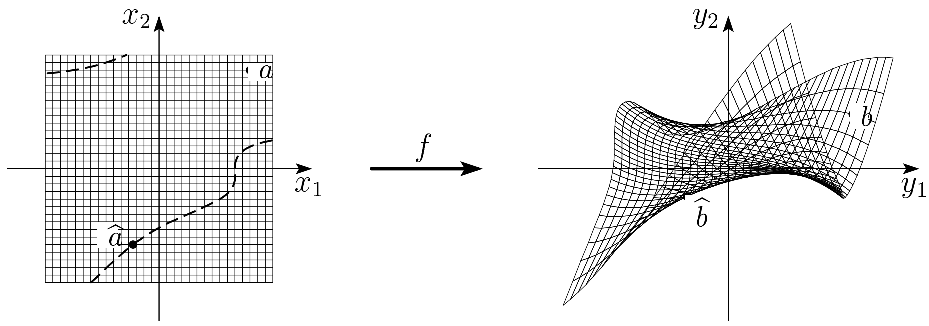


Figure III.4: Illustration du théorème d'inversion locale

Pour le point $a = (1.2, 1.3)^t$, nous avons dessiné son image $b = f(a)$. Comme la matrice Jacobienne

$$f'(x) = \begin{pmatrix} 1 & 0.6\, x_2^2 \\ 0.2\, x_2^4 - 0.3\, x_2 - 0.3\, x_1^2 & 0.8\,(x_1 + 0.5)\, x_2^3 - 0.3\,(x_1 - 1) \end{pmatrix}$$

est inversible en $x = a$, le théorème d'inversion locale nous permet de conclure que, pour y suffisamment proche de b, le système $f(x) = y$ possède une solution unique proche de a. Ce résultat d'existence et d'unicité d'une solution nous encourage à utiliser des méthodes numériques.

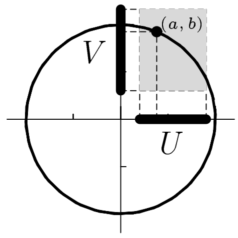
Les courbes traitillées dans le plan (x_1, x_2) de la figure III.4 indiquent les points où $f'(x)$ n'est pas inversible, c.-à-d. où $\det f'(x) = 0$. On ne peut donc pas appliquer le théorème d'inversion locale en ces points. En effet, pour le point $\widehat{b}$ qui est l'image du point $\widehat{a}$, il n'existe pas de voisinage V tel que, pour tout $y \in V$, le système $f(x) = y$ possède une solution unique. Le dessin nous montre qu'il y a des points y proche de $\widehat{b}$, pour lesquels $f(x) = y$ ne possède pas de solution dans le carré considéré, et d'autres, pour lesquels il y a deux solutions proche de $\widehat{a}$.

Théorème des fonctions implicites. Pour une fonction $g : \mathbb{R}^k \times \mathbb{R}^m \to \mathbb{R}^m$, considérons l'équation $g(x, y) = 0$, c.-à-d.

$$\begin{aligned} g_1(x_1, \ldots, x_k, y_1, \ldots, y_m) &= 0 \\ g_2(x_1, \ldots, x_k, y_1, \ldots, y_m) &= 0 \\ &\cdots \\ g_m(x_1, \ldots, x_k, y_1, \ldots, y_m) &= 0\,. \end{aligned} \tag{6.3}$$

Nous allons étudier si, pour un $x \in \mathbb{R}^k$ donné, cette équation possède une solution $y \in \mathbb{R}^m$, et si cette solution est localement unique.

Afin de mieux préciser le problème posé, considérons un exemple simple. Pour la fonction $g(x, y) = x^2 + y^2 - 1$, l'ensemble des points satisfaisant $g(x, y) = 0$ est un cercle. Fixons maintenant un point (a, b) sur ce cercle. Nous voyons sur le dessin d'à côté qu'il existe des voisinages U et V de a respectivement de b tels que pour $x \in U$ il existe un unique $y \in V$ avec $g(x, y) = 0$. Cette affirmation est vraie tant que $b \neq 0$, mais elle est fausse si $b = 0$. Nous constatons que la condition $b = 0$ est équivalente à $\frac{\partial g}{\partial y}(a, b) = 0$.



Nous allons généraliser ce résultat à des fonctions $g(x, y)$ plus compliquées et à la situation où x et y sont des vecteurs comme dans (6.3).

Théorème 6.4 (théorème des fonctions implicites) *Soit $g : \mathbb{R}^k \times \mathbb{R}^m \to \mathbb{R}^m$ continûment différentiable et supposons qu'au point $(a, b) \in \mathbb{R}^k \times \mathbb{R}^m$*

$$g(a, b) = 0 \qquad \textit{et} \qquad \frac{\partial g}{\partial y}(a, b) \quad \textit{est inversible.}$$

Il existe alors un voisinage U de a, un voisinage V de b et une application $\varphi : U \to V$ (continûment différentiable) tels que pour $(x, y) \in U \times V$ on a

$$g(x, y) = 0 \qquad \Longleftrightarrow \qquad y = \varphi(x). \tag{6.4}$$

Remarquons que, pour la fonction g de (6.3), la dérivée partielle $\frac{\partial g}{\partial y}$ représente la matrice carrée

$$\frac{\partial g}{\partial y}(x, y) = \begin{pmatrix} \frac{\partial g_1}{\partial y_1}(x, y) & \cdots & \frac{\partial g_1}{\partial y_m}(x, y) \\ \vdots & & \vdots \\ \frac{\partial g_m}{\partial y_1}(x, y) & \cdots & \frac{\partial g_m}{\partial y_m}(x, y) \end{pmatrix}.$$

Pour pouvoir appliquer ce théorème, il faut que $\frac{\partial g}{\partial y}$ soit inversible, i.e. $\det \frac{\partial g}{\partial y}(a, b) \neq 0$. Il ne suffit pas que cette matrice soit non nulle.

Démonstration. Considérons l'application

$$F : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} x \\ g(x, y) \end{pmatrix} \qquad \text{avec} \qquad F'(a, b) = \begin{pmatrix} I & 0 \\ \frac{\partial g}{\partial x}(a, b) & \frac{\partial g}{\partial y}(a, b) \end{pmatrix}.$$

L'hypothèse sur l'inversibilité de $\frac{\partial g}{\partial y}(a, b)$ implique que $F'(a, b)$ est aussi inversible. Le théorème d'inversion locale montre alors que F est un difféomorphisme local près de (a, b). L'application inverse de F est nécessairement de la forme

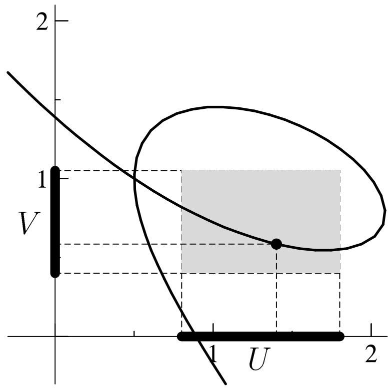
$$F^{-1} : \begin{pmatrix} u \\ v \end{pmatrix} \mapsto \begin{pmatrix} u \\ h(u, v) \end{pmatrix}.$$

Comme $F \circ F^{-1}$ est l'identité, on a $g(u, h(u, v)) = v$ et donc $g(u, h(u, 0)) = 0$. L'affirmation du théorème suit avec $\varphi(x) := h(x, 0)$. Comme F est un difféomorphisme local près de (a, b), il n'y a pas d'autres solutions de $g(x, y) = 0$ que celle de la forme $(x, \varphi(x))$. $\square$

Exemple 6.5 Considérons la fonction

$$g(x,y) = 16x^3 - 84x^2 + 162x - 89 + 27y^3 + 54xy^2 - 108y^2 + 36x^2y - 180xy + 162y,$$

pour laquelle la condition $g(x,y)=0$ donne une courbe nettement plus compliquée que le cercle de la discussion d'avant le théorème. Pour le point $(a,b) \approx (1.4, 0.5853)$ sur cette courbe (i.e. $g(a,b)=0$), on vérifie par un calcul direct que $\frac{\partial g}{\partial y}(a,b) \approx -3.702 \neq 0$. Le théorème des fonctions implicites implique qu'il existe des voisinages U de a, V de b et une fonction différentiable $\varphi : U \to V$ tels que (6.4) est vraie pour $(x,y) \in U \times V$ (voir la figure).



Les points de la courbe ayant une tangente verticale (resp. horizontale) peuvent être trouvés par la condition $\frac{\partial g}{\partial y}(x,y)=0$ (resp. $\frac{\partial g}{\partial x}(x,y)=0$). Au point du croisement, on a nécessairement $g(x,y)=0$, $\frac{\partial g}{\partial y}(x,y)=0$ et $\frac{\partial g}{\partial x}(x,y)=0$ (trois conditions pour deux inconnues).

III.7 Surfaces et sous-variétés

Dans ce paragraphe, nous discutons deux formulations mathématiques différentes de surfaces, et nous généralisons les notions de courbes et de surfaces à des objets géométriques (sous-variétés) de dimension plus grande.

Exemples de surfaces dans $\mathbb{R}^3$. Avant de donner une définition rigoureuse et de discuter l'existence des paramétrisations, considérons quelques exemples simples.

Exemple 7.1 (sphère dans $\mathbb{R}^3$) La sphère dans $\mathbb{R}^3$ de rayon 1, centrée à l'origine, est donnée par

$$\mathcal{M} = \{(x_1,x_2,x_3) \mid x_1^2 + x_2^2 + x_3^2 - 1 = 0\}.$$

La sphère, ou une partie de celle-ci, peut aussi être décrite comme l'image d'une des applications

$$\varphi(x_1,x_2) = \begin{pmatrix} x_1 \\ x_2 \\ \sqrt{1-x_1^2-x_2^2} \end{pmatrix} \qquad \text{ou} \qquad \chi(\alpha,\theta) = \begin{pmatrix} \cos\alpha\sin\theta \\ \sin\alpha\sin\theta \\ \cos\theta \end{pmatrix}$$

(représentation paramétrique de la sphère). Les images des droites $\alpha = Const$ et $\theta = Const$ sont dessinées dans la figure III.5 (gauche).

Exemple 7.2 (surface cylindrique) La surface cylindrique peut être décrite par

$$\mathcal{M} = \{(x_1,x_2,x_3) \mid x_1^2 + x_2^2 - 1 = 0\}$$

(voir le dessin du milieu de la figure III.5) ou par une paramétrisation

$$\varphi(\alpha,x_3) = (\cos\alpha,\, \sin\alpha,\, x_3)^t.$$

Pour la représentation graphique, on a de nouveau dessiné les images des droite $\alpha = Const$ et $x_3 = Const$, ce qui montre l'utilité des représentations paramétriques.

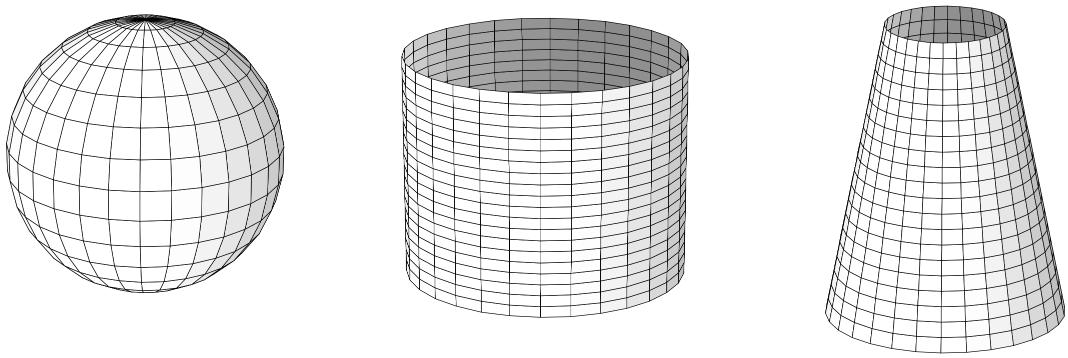


Figure III.5: Surfaces dans $\mathbb{R}^3$: sphère, partie d'un cylindre et d'un cône

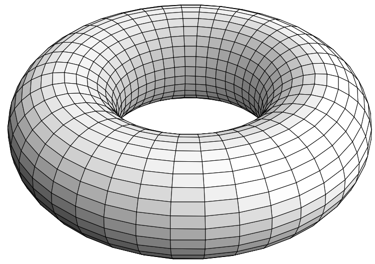
Exemple 7.3 (cône) Le troisième objet de la figure III.5 est une partie du cône

$$\mathcal{M} = \{(x_1, x_2, x_3) \,|\, x_1^2 + x_2^2 - (1 - x_3)^2 = 0\}.$$

Une représentation paramétrique est

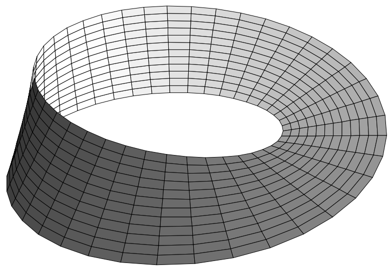
$$\varphi(\alpha, x_3) = ((1 - x_3)\cos\alpha,\ (1 - x_3)\sin\alpha,\ x_3)^t.$$

Exemple 7.4 (tore de révolution) Dans le plan (x_1, x_3), considérons le cercle centré en $(d, 0)$ de rayon ρ ($0 < \rho < d$) et tournons-le autour de l'axe x_3. Ceci donne une surface de révolution avec la paramétrisation



$$\varphi(\alpha, \beta) = \begin{pmatrix} (d + \rho\cos\alpha)\cos\beta \\ (d + \rho\cos\alpha)\sin\beta \\ \rho\sin\alpha \end{pmatrix}.$$

Exemple 7.5 (ruban de Möbius) Considérons une tige de longueur 2 (paramétrisée par $-1 < t < 1$) et tournons-la autour de son centre et, en même temps, deux fois plus vite autour d'un axe à distance d. Ceci donne la paramétrisation



$$\varphi(t, \alpha) = \begin{pmatrix} (d + t\cos\alpha)\cos 2\alpha \\ (d + t\cos\alpha)\sin 2\alpha \\ t\sin\alpha \end{pmatrix}.$$

Après tant d'exemples, nous sommes arrivés au point où une définition rigoureuse d'une surface s'avère utile. Ceci nous permet en même temps d'introduire des objets géométriques d'une dimension plus grande.

Définition 7.6 Un ensemble $\mathcal{M} \neq \emptyset$ est *une sous-variété de* $\mathbb{R}^n$ si, pour tout $a \in \mathcal{M}$, il existe un voisinage U de a ($U \subset \mathbb{R}^n$) et une application différentiable $g : U \to \mathbb{R}^{n-k}$ avec $g(a) = 0$ et $g'(a)$ de rang maximal $n - k$ tels que

$$\mathcal{M} \cap U = \{x \in U \mid g(x) = 0\};$$

on appelle k la *dimension de la sous-variété.*

Théorème 7.7 (existence d'une paramétrisation locale) *Soient $\mathcal{M}$ une sous-variété de $\mathbb{R}^n$ de dimension k, et $a \in \mathcal{M}$. Alors il existe un voisinage $U \subset \mathbb{R}^n$ de a, un voisinage $V \subset \mathbb{R}^k$ d'un point $z_0 \in \mathbb{R}^k$ et une application différentiable $\varphi : V \to U$ avec $\varphi(z_0) = a$ et $\varphi'(z_0)$ de rang k tels que*

$$\mathcal{M} \cap U = \varphi(V) = \{\varphi(z) \mid z \in V\}.$$

Démonstration. Proche du point $a \in \mathcal{M}$, la sous-variété est définie par la condition $g(x) = 0$, et on sait que le rang de la matrice Jacobienne $g'(a)$ est maximal et donc égal à $n - k$. Après une permutation des x_i (si nécessaire), on peut supposer que

$$\begin{pmatrix} \frac{\partial g_1}{\partial x_{k+1}}(a) & \cdots & \frac{\partial g_1}{\partial x_n}(a) \\ \vdots & & \vdots \\ \frac{\partial g_{n-k}}{\partial x_{k+1}}(a) & \cdots & \frac{\partial g_{n-k}}{\partial x_n}(a) \end{pmatrix} \qquad \text{est inversible.}$$

Par le théorème des fonctions implicites, on peut donc localement exprimer $x_{k+1}, \ldots, x_n$ en fonction de $x_1, \ldots, x_k$. Avec la notation $z := (x_1, \ldots, x_k)^t$, $z_0 := (a_1, \ldots, a_k)^t$, et $y := (x_{k+1}, \ldots, x_n)^t$, ceci implique qu'il existe une application différentiable $\widehat{\varphi}(z)$ tel que, proche de a, $g(x) = g(z, y) = 0$ est équivalent à $y = \widehat{\varphi}(z)$. L'affirmation du théorème suit avec $\varphi(z) := (z, \widehat{\varphi}(z))$. □

Ce théorème explique le terme "dimension k", parce que la sous-variété est décrite à l'aide de k paramètres car $z = (z_1, \ldots, z_k)^t \in \mathbb{R}^k$. Des cas particuliers de sous-variétés de $\mathbb{R}^n$ sont:

- $k = 1$ et n arbitraire : *courbe* dans $\mathbb{R}^n$. Elle peut localement être décrite par une fonction $\varphi : \mathbb{R} \to \mathbb{R}^n$ comme étant l'image d'un intervalle ouvert, ou par une application $g : \mathbb{R}^n \to \mathbb{R}^{n-1}$ comme étant $\{x \in U \mid g(x) = 0\}$.

- $k = 2$ et $n = 3$: *surface* dans $\mathbb{R}^3$. Elle est décrite soit par $\varphi : \mathbb{R}^2 \to \mathbb{R}^3$, soit par $g : \mathbb{R}^3 \to \mathbb{R}$ (voir les exemples 7.1 à 7.5).

- $k = n - 1$ et n arbitraire : *hypersurface*. Elle est décrite soit par $\varphi : \mathbb{R}^{n-1} \to \mathbb{R}^n$, soit par $g : \mathbb{R}^n \to \mathbb{R}$.

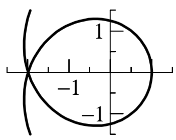
Pour mieux comprendre la définition d'une sous-variété, donnons quelques contre-exemples. L'ensemble

$$\mathcal{M} = \{(x, y) \mid xy = 0\}$$

n'est pas une sous-variété près de $(0, 0)$ car, avec $g(x, y) = xy$, la dérivée $g'(0, 0) = (0, 0)$ n'est pas de rang maximal. En effet, proche de l'origine, $\mathcal{M}$ ne peut pas être décrit comme étant l'image d'une fonction différentiable. Par contre, l'ensemble $\mathcal{M} \setminus \{(0, 0)\}$ est une sous-variété de $\mathbb{R}^2$.

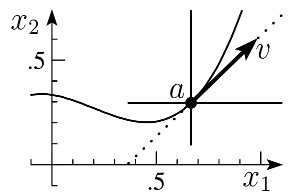
Similairement, l'ensemble $\mathcal{M} = \{\varphi(t) \mid t \in \mathbb{R}\}$ avec

$$\varphi(t) = (1 + 0.1\, t^2) \begin{pmatrix} \cos t \\ \sin t \end{pmatrix}$$



n'est pas une sous-variété près de $a = (-1 - 0.1\, \pi^2,\, 0)$. En effet, il n'existe pas de $g : \mathbb{R}^2 \to \mathbb{R}$ avec $g(a) = 0$ et $g'(a)$ de rang 1, car sinon il y aurait une contradiction avec le théorème des fonctions implicites. Ce contre-exemple montre que l'image d'un intervalle par $\varphi : \mathbb{R} \to \mathbb{R}^n$ n'est pas toujours une sous-variété, même si φ est continûment différentiable avec $\varphi'(t) \neq 0$.

III.8 Espace tangent



Courbes. Pour une courbe $\gamma : I \to \mathbb{R}^n$ (différentiable et $\dot\gamma(0) \neq 0$), la tangente est la droite donnée par $\tau(t) = a + tv$ où $a = \gamma(0)$ et $v = \dot\gamma(0)$. En déplaçant l'origine au point a, la tangente en a à la courbe $\mathcal{M} = \gamma(I)$ devient

$$T_a\mathcal{M} = \{tv \mid t \in \mathbb{R}\}$$

qui est un espace vectoriel. Dans les variables originelles, la tangente est alors l'espace affine $a + T_a\mathcal{M}$ qui est un sous-ensemble de $\mathbb{R}^n$.

Surfaces dans $\mathbb{R}^3$. Comme exemple, prenons l'ellipsoïde

$$\mathcal{M} = \Big\{(x,y,z) \;\Big|\; \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1\Big\} \quad \text{avec paramétrisation} \quad \varphi(\alpha,\theta) = \begin{pmatrix} a\,\cos\alpha\,\sin\theta \\ b\,\sin\alpha\,\sin\theta \\ c\,\cos\theta \end{pmatrix}.$$

Pour déterminer le plan tangent à un point $a = (x_0, y_0, z_0) = \varphi(\alpha_0, \theta_0) \in \mathcal{M}$, nous considérons les courbes paramétriques $\gamma(t) = \varphi(t, \theta_0)$ et $\delta(t) = \varphi(\alpha_0, t)$ dessinées dans la figure III.6. Le dessin de gauche montre en plus les tangentes (en gris) $\tau(t) = a + tv_1$ avec $v_1 = \dot\gamma(\alpha_0)$ et $\sigma(t) = a + tv_2$ avec $v_2 = \dot\delta(\theta_0)$. Les deux vecteurs v_1 et v_2 engendrent le plan tangent à l'ellipsoïde. Il est donné par $a + T_a\mathcal{M}$ avec

$$T_a\mathcal{M} = \{t_1v_1 + t_2v_2 \mid t_1, t_2 \in \mathbb{R}\} \qquad \text{où} \qquad v_1 = \frac{\partial\varphi}{\partial\alpha}(\alpha_0,\theta_0), \quad v_2 = \frac{\partial\varphi}{\partial\theta}(\alpha_0,\theta_0).$$

Pour d'autres courbes dans $\mathcal{M}$ passant par a, la tangente est aussi dans $a + T_a\mathcal{M}$ (voir le dessin de droite de la figure III.6).

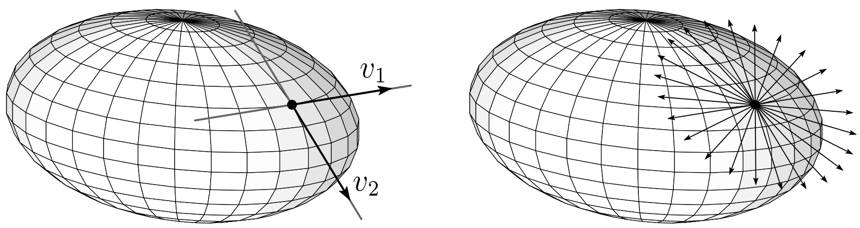


Figure III.6: Illustration de la définition de l'espace tangent

Définition 8.1 Soient $\mathcal{M} \subset \mathbb{R}^n$ une sous-variété de $\mathbb{R}^n$ et $a \in \mathcal{M}$. *L'espace tangent à* $\mathcal{M}$ *en* a est

$$T_a\mathcal{M} = \left\{ v \in \mathbb{R}^n \;\middle|\; \begin{array}{c} \text{il existe } \gamma : (-\varepsilon,\varepsilon) \to \mathbb{R}^n \text{ continûment différentiable tel que} \\ \gamma(t) \in \mathcal{M} \text{ pour } t \in (-\varepsilon,\varepsilon) \text{ et } \gamma(0) = a \text{ et } \dot\gamma(0) = v \end{array} \right\}.$$

Cette définition donne une jolie interprétation géométrique de l'espace tangent. Pour un calcul explicite de $T_a\mathcal{M}$, les formules du théorème suivant sont plus utiles.

Théorème 8.2 *Soient* $\mathcal{M} \subset \mathbb{R}^n$ *une sous-variété de dimension* k *et* $a \in \mathcal{M}$.

- *Si, proche de* a, $\mathcal{M}$ *est donné par une paramétrisation locale* $\varphi : V \to \mathbb{R}^n$, *c.-à-d. on a* $\mathcal{M} \cap U = \{\varphi(z) \mid z \in V\}$ *où* $\varphi(z_0) = a$ *avec* $z_0 \in V \subset \mathbb{R}^k$, *alors*

$$T_a\mathcal{M} = \operatorname{Im}\varphi'(z_0) = \{\varphi'(z_0)\,t \mid t \in \mathbb{R}^k\}. \tag{8.1}$$

- *Si $\mathcal{M}$ est localement donné par $\mathcal{M} \cap U = \{x \in U \mid g(x) = 0\}$, alors*

$$T_a\mathcal{M} = \ker\, g'(a) = \{v \in \mathbb{R}^n \mid g'(a)v = 0\}. \tag{8.2}$$

Démonstration. Soit $\delta(s)$ une courbe dans $\mathbb{R}^k$ avec $\delta(0) = z_0$ et $\dot{\delta}(0) = t$ (par exemple la droite $\delta(s) = z_0 + st$). Alors $\gamma(s) := \varphi(\delta(s))$ est une courbe dans $\mathcal{M}$ satisfaisant $\gamma(0) = \varphi(z_0) = a$ et $\dot{\gamma}(0) = \varphi'(z_0)\,\dot{\delta}(0) = \varphi'(z_0)\,t$. Ceci implique $\mathrm{Im}\,\varphi'(z_0) \subset T_a\mathcal{M}$.

Si $\gamma(t)$ est une courbe dans $\mathcal{M}$ satisfaisant $\gamma(0) = a$ et $\dot{\gamma}(0) = v$, alors $g(\gamma(t)) = 0$ et $g'(a)\dot{\gamma}(0) = 0$. Ceci implique $T_a\mathcal{M} \subset \ker\, g'(a)$.

Par définition de la sous-variété $\mathcal{M}$, les deux espaces vectoriels $\mathrm{Im}\,\varphi'(z_0)$ et $\ker\, g'(a)$ ont la même dimension k. On déduit donc des inclusions $\mathrm{Im}\,\varphi'(z_0) \subset T_a\mathcal{M} \subset \ker\, g'(a)$ les égalités $\mathrm{Im}\,\varphi'(z_0) = T_a\mathcal{M} = \ker\, g'(a)$. □

Ce théorème montre que l'espace tangent $T_a\mathcal{M}$ est un espace vectoriel de dimension k (même dimension que la sous-variété). La formule (8.1) signifie que $T_a\mathcal{M}$ est engendré par les vecteurs

$$\frac{\partial\varphi}{\partial z_1}(z_0), \ldots, \frac{\partial\varphi}{\partial z_k}(z_0)$$

(voir les exemples qui précédent la définition 8.1). La formule (8.2) peut aussi être trouvée de la manière suivante. Considérons, par exemple, l'ellipsoïde donnée par

$$g(x, y, z) = \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} - 1 = 0.$$

Proche d'un point $(x_0, y_0, z_0) \in \mathcal{M}$, la différentiabilité de g implique que

$$g(x, y, z) = g(x_0, y_0, z_0) + \frac{2x_0}{a^2}(x - x_0) + \frac{2y_0}{b^2}(y - y_0) + \frac{2z_0}{c^2}(z - z_0) + \ldots .$$

En négligeant les termes supérieurs (les trois points), on obtient la meilleure approximation linéaire de $g(x, y, z)$ proche de (x_0, y_0, z_0). Le plan tangent est donc

$$\Big\{(x, y, z) \;\Big|\; \frac{2x_0}{a^2}(x - x_0) + \frac{2y_0}{b^2}(y - y_0) + \frac{2z_0}{c^2}(z - z_0) = 0\Big\}.$$

Cela correspond en effet à la formule (8.2).

III.9 Exercices

Chapter IV

Optimisation

Ce chapitre est consacré à l'étude des maxima et minima d'une fonction à plusieurs variables. On dit que $f : K \to \mathbb{R}$ (avec $K \subset \mathbb{R}^n$) possède un *minimum local ou relatif* au point $a \in K$ si

$$f(x) \geq f(a) \qquad \text{pour tout } x \in K \cap U \text{ où } U \text{ est un voisinage de } a.$$

Elle possède un *minimum global* au point $a \in K$ si

$$f(x) \geq f(a) \qquad \text{pour tout } x \in K.$$

On parle d'un minimum *strict* si dans ces définitions $f(x) > f(a)$ pour $x \neq a$. Pour obtenir des résultats concernant les maxima, il suffit de remplacer f par $-f$.

Nous allons étudier des minima relatifs dans les cas où $K = \mathbb{R}^n$ ou une sous-variété de $\mathbb{R}^n$. Nous discuterons aussi l'algorithme du simplexe pour le cas où f est une fonction linéaire et K un polyèdre.

Remarquons que le théorème 4.5 du chapitre I nous dit que, si $f : K \to \mathbb{R}$ est continue et si K est compact, il existe toujours un minimum global. Malheureusement, sa démonstration n'est pas constructive.

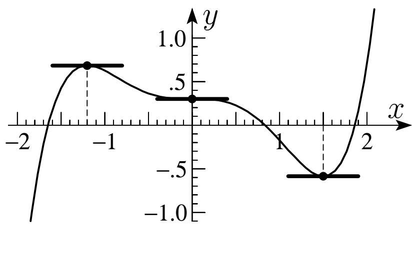
IV.1 Minima relatifs

Rappelons que, pour une fonction différentiable $f : \mathbb{R} \to \mathbb{R}$, la condition $f'(a) = 0$ est nécessaire pour un extremum. Si la fonction est deux fois continûment différentiable, on a

$$\left\{ \begin{array}{l} f'(a) = 0 \\ f''(a) > 0 \end{array} \right\} \quad \Rightarrow \quad \{ \text{ minimum relatif en } a \} \quad \Rightarrow \quad \left\{ \begin{array}{l} f'(a) = 0 \\ f''(a) \geq 0 \end{array} \right\} \tag{1.1}$$

Considérons, par exemple, la fonction

$$f(x) = 0.2\,x^5 - 0.075\,x^4 - 0.6\,x^3 + 0.3.$$



Sa dérivée est $f'(x) = x^2(x+1.2)(x-1.5)$ et s'annule pour $a = -1.2$, $a = 0$ et $a = 1.5$. En $a = -1.2$ la fonction f possède un maximum relatif ($f''(-1.2) = -3.888 < 0$), en $a = 1.5$ un minimum relatif ($f''(1.5) = 3.825 > 0$) et au point $a = 0$ on n'a ni un maximum ni un minimum ($f''(0) = 0$).

Le but de ce paragraphe est de généraliser l'affirmation (1.1) à des fonctions à plusieurs variables. Comme (1.1) se démontre à l'aide de la formule de Taylor, nous rappelons brièvement les premiers termes de la série de Taylor pour une fonction à plusieurs variables.

Pour une fonction $f(x) = f(x_1, \ldots, x_n)$ qui est deux fois continûment différentiable, la série de Taylor avec reste sous forme intégrale (voir aussi (III.4.10)) est

$$\begin{aligned} f(a+v) &= f(a) + f'(a)\,v + \int_0^1 (1-t) f''(a+tv)(v,v)\,dt \\ &= f(a) + f'(a)\,v + \frac{1}{2} f''(a)(v,v) + \int_0^1 (1-t)\Big(f''(a+tv) - f''(a)\Big)(v,v)\,dt. \end{aligned}$$

En utilisant le fait que $\|(f''(a+tv) - f''(a))(v,v)\| \le \|f''(a+tv) - f''(a)\| \cdot \|v\|^2$ et que toutes les deuxièmes dérivées partielles sont continues en a, nous obtenons

$$f(a+v) = f(a) + f'(a)\,v + \frac{1}{2} f''(a)(v,v) + r(v)\|v\|^2$$

où $r(v) \to 0$ si $v \to 0$. Comme $f(x)$ est une fonction scalaire, cette formule peut aussi être écrite sous la forme

$$f(a+v) = f(a) + \nabla f(a)^t v + \frac{1}{2} v^t H(a)\,v + r(v)\|v\|^2 \tag{1.2}$$

où

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix} \qquad \text{et} \qquad H(x) = \nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \ldots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x) & \ldots & \frac{\partial^2 f}{\partial x_n^2}(x) \end{pmatrix}.$$

Le vecteur $\nabla f(x)$ est le *gradient* de la fonction $f(x)$ et $H(x)$ est sa *matrice Hessienne*. Cette matrice est symétrique car les deuxièmes dérivées partielles ne dépendent pas de l'ordre de la différentiation. C'est surtout la formule (1.2) qu'on va utiliser pour la suite.

La généralisation de l'affirmation (1.1) est la suivante:

Théorème 1.1 *Soient $f : \mathbb{R}^n \to \mathbb{R}$ deux fois continûment différentiable et $a \in \mathbb{R}^n$. Alors on a*

$$\left\{ \begin{array}{c} \nabla f(a) = 0 \\ H(a) \text{ est définie positive} \end{array} \right\} \Rightarrow \left\{ \begin{array}{c} f \text{ possède un} \\ \text{minimum relatif en } a \end{array} \right\} \Rightarrow \left\{ \begin{array}{c} \nabla f(a) = 0 \\ H(a) \text{ est semi-définie positive} \end{array} \right\}$$

Démonstration. Rappelons que si $\lambda_{\min}$ est la plus petite valeur propre d'une matrice symétrique H, on a

$$v^t H v \ge \lambda_{\min}\, v^t v \quad \text{pour } v \in \mathbb{R}^n. \tag{1.3}$$

Ceci est une conséquence immédiate du corollaire II.2.6.

Pour $\nabla f(a) = 0$, la formule de Taylor (1.2) et l'estimation (1.3) impliquent que

$$f(a+v) - f(a) \ge (\lambda_{\min} + r(v))\,\|v\|^2.$$

Si $\lambda_{\min} > 0$ (c.-à-d. si la matrice $H(a)$ est définie positive), on a $f(a+v) > f(a)$ pour $v \ne 0$ suffisamment petit. On a même démontré que a est un minimum strict.

La condition nécessaire de l'affirmation du théorème est une conséquence de (1.1) appliqué à la fonction $g(t) = f(a+tv)$. □

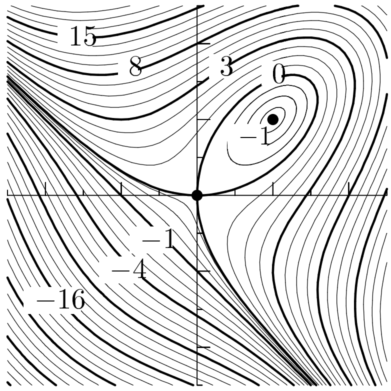
Corollaire 1.2 *Situation comme dans le théorème 1.1. Alors on a*

$$\left\{ \begin{array}{c} \nabla f(a) = 0 \\ -H(a) \text{ est définie positive} \end{array} \right\} \Rightarrow \left\{ \begin{array}{c} f \text{ possède un} \\ \text{maximum relatif en } a \end{array} \right\} \Rightarrow \left\{ \begin{array}{c} \nabla f(a) = 0 \\ -H(a) \text{ est semi-définie positive} \end{array} \right\}$$

Démonstration. Cette affirmation est une conséquence immédiate du théorème 1.1, car f possède un maximum en a si et seulement si $-f$ possède un minimum en a. □

Exemple 1.3 Cherchons les minima relatifs de la fonction

$$f(x_1, x_2) = x_1^3 + x_2^3 - 3x_1x_2$$



(folium cartesii). Ses courbes de niveau sont dessinées dans la figure d'à côté. On commence par calculer le gradient

$$\nabla f(x) = \begin{pmatrix} 3x_1^2 - 3x_2 \\ 3x_2^2 - 3x_1 \end{pmatrix}$$

et les valeurs de a où $\nabla f(a) = 0$. Pour cet exemple simple, on trouve tout de suite $a = (0, 0)$ et $a = (1, 1)$. Pour des fonctions plus compliquées, on est obligé d'utiliser des méthodes numériques.

Pour vérifier s'il s'agit d'un minimum, nous calculons la matrice Hessienne

$$H(x) = \begin{pmatrix} 6x_1 & -3 \\ -3 & 6x_2 \end{pmatrix}.$$

Le valeurs propres de $H(a)$ sont les zéros du polynôme caractéristique $(\lambda - 6a_1)(\lambda - 6a_2) - 9$. Au point $a = (0, 0)$, nous trouvons $\lambda_1 = -3$ et $\lambda_2 = 3$ comme valeurs propres de $H(a)$. Il ne s'agit donc ni d'un minimum ni d'un maximum, comme on peut le voir sur le dessin. Par contre, les valeurs propres de $H(a)$ pour $a = (1, 1)$ sont $\lambda_1 = 3$ et $\lambda_2 = 9$. Comme elles sont positives, il s'agit d'un minimum relatif.

Remarquons que, pour une fonction $f : \mathbb{R}^n \to \mathbb{R}$, la condition $\nabla f(x) = 0$ donne n équations à n inconnues à résoudre, notamment

$$\frac{\partial f}{\partial x_1}(x_1, \ldots, x_n) = \frac{\partial f}{\partial x_2}(x_1, \ldots, x_n) = \ldots = \frac{\partial f}{\partial x_n}(x_1, \ldots, x_n) = 0.$$

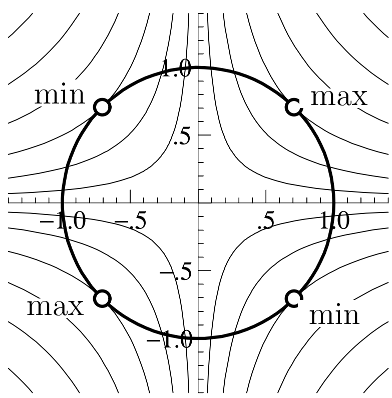
Pour vérifier si une solution $a \in \mathbb{R}^n$ représente un minimum relatif de $f(x)$, il faut calculer les valeurs propres de la matrice Hessienne $H(a)$. Les deux problèmes (la résolution d'un système non linéaire dans $\mathbb{R}^n$ et le calcul des valeurs propres d'une matrice de dimension n) seront traités dans le cours 'Analyse Numérique'.

IV.2 Minima conditionnels – multiplicateurs de Lagrange

Pour mieux comprendre les idées, commençons par le problème

$$f(x_1, x_2) = x_1x_2 \to \min$$

$$\text{sur} \quad \mathcal{M} = \{(x_1, x_2) \mid g(x_1, x_2) := x_1^2 + x_2^2 - 1 = 0\}.$$



On voit sur le dessin d'à côté que les solutions de ce problème sont $(\frac{-\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$ et $(\frac{\sqrt{2}}{2}, \frac{-\sqrt{2}}{2})$.

Raisonnement géométrique. On sait que $\nabla f(a)$ est orthogonal aux lignes de niveau car $f(a+v) - f(a) = \nabla f(a)^t v + \mathcal{O}(\|v\|^2)$. De plus, $\nabla g(a)$ est orthogonal à $\mathcal{M}$ (formule (8.2) du paragraphe III.8). Si la courbe de niveau de $f(x)$ et la sous-variété $\mathcal{M}$ se coupent transversalement en a, on ne peut pas avoir un minimum. Donc, aux solutions de ce problème, on a

$$\nabla f(a) = \lambda \, \nabla g(a).$$

Raisonnement analytique. La sous-variété $\mathcal{M} = \{(x_1, x_2) \mid x_1^2+x_2^2-1 = 0\}$ peut être paramétrisée à l'aide de $\varphi(t) = (\cos t, \sin t)^t$. Le problème qui consiste à minimiser $f(x)$ sur $\mathcal{M}$ est alors équivalent à minimiser la fonction $F(t) = f(\varphi(t)) = \cos t \sin t = \frac{1}{2}\sin 2t$ sur $\mathbb{R}$. La condition $F'(t) = \cos 2t = 0$ est satisfaite pour $t = \pi/4,\ 3\pi/4,\ 5\pi/4,$ etc. Comme $F''(t) = -2\sin 2t$, les minima sont obtenus pour $t = 3\pi/4$ et pour $t = 7\pi/4$.

Pour une fonction arbitraire $f(x_1, x_2)$ et pour une sous-variété de $\mathbb{R}^2$ donnée par une courbe paramétrique $\varphi(t)$, l'approche analytique donne pour $F(t) = f(\varphi(t))$ la condition

$$F'(t) = f'(\varphi(t))\,\dot\varphi(t) = \nabla f(\varphi(t))^t \dot\varphi(t) = 0.$$

Avec l'interprétation géométrique, on a donc au minimum:

$$\nabla f(a) \ \perp \ \text{tangente à } \mathcal{M} \text{ en } a \quad \Leftrightarrow \quad \nabla f(a) \parallel \nabla g(a) \quad \Leftrightarrow \quad \nabla f(a) = \lambda \nabla g(a).$$

Le but est de généraliser cet exemple à une fonction f de n variables et à une sous-variété arbitraire de $\mathbb{R}^n$ de dimension k.

Le problème à traiter dans ce paragraphe est alors le suivant: soient données une fonction $f : \mathbb{R}^n \to \mathbb{R}$ continûment différentiable et une sous-variété $\mathcal{M}$ de dimension k. Calculer les solutions de

$$f(x) \to \min \quad \text{sur } \mathcal{M}, \tag{2.1}$$

c.-à-d. on cherche $a \in \mathcal{M}$ tel que $f(x) \ge f(a)$ pour $x \in \mathcal{M}$ proche de a. On dit aussi que $a \in \mathcal{M}$ est un minimum relatif de la fonction $f|_{\mathcal{M}}$ (restriction de f à $\mathcal{M}$).

Théorème 2.1 *Soit $f : \mathbb{R}^n \to \mathbb{R}$ deux fois continûment différentiable et soit $\varphi : \mathbb{R}^k \to \mathbb{R}^n$ une paramétrisation locale de $\mathcal{M}$ près de $a \in \mathcal{M}$ (et $\varphi(z_0) = a$). Alors*

$$\left\{ \begin{array}{c} \nabla f(a)^t \varphi'(z_0) = 0 \textit{ et} \\ H \textit{ est définie positive} \end{array} \right\} \Rightarrow \left\{ \begin{array}{c} f|_{\mathcal{M}} \textit{ possède un} \\ \textit{minimum relatif en } a \end{array} \right\} \Rightarrow \left\{ \begin{array}{c} \nabla f(a)^t \varphi'(z_0) = 0 \textit{ et} \\ H \textit{ est semi-définie positive} \end{array} \right\}$$

où H est la matrice Hessienne de l'application $z \mapsto f(\varphi(z))$, c.-à-d.

$$H := \varphi'(z_0)^t \nabla^2 f(a)\, \varphi'(z_0) + \sum_{l=1}^n \frac{\partial f}{\partial x_l}(a) \nabla^2 \varphi_l(z_0). \tag{2.2}$$

Notons que la matrice Hessienne $\nabla^2 f(a)$ de la fonction f est de dimension n, mais que la matrice H est seulement de dimension k.

Démonstration. La restriction de la fonction f sur $\mathcal{M}$ possède un minimum relatif en $a \in \mathcal{M}$ si et seulement si $F(z) := f(\varphi(z))$ en possède un en z_0. En observant que $F'(z) = f'(\varphi(z))\varphi'(z) = \nabla f(\varphi(z))^t \varphi'(z)$ et que $\nabla^2 F(z) = \varphi'(z)^t \nabla^2 f(\varphi(z))\varphi'(z) + \sum_{l=1}^n \frac{\partial f}{\partial x_l}(\varphi(z))\nabla^2 \varphi_l(z)$, l'affirmation est une conséquence du théorème 1.1. □

Si on veut éviter la notation matricielle, la condition $F'(z_0) = \nabla f(a)^t \varphi'(z_0) = 0$ s'écrit

$$\frac{\partial F}{\partial z_i}(z_0) = \sum_{l=1}^n \frac{\partial f}{\partial x_l}(a)\, \frac{\partial \varphi_l}{\partial z_i}(z_0) = 0 \qquad \text{pour} \quad i = 1, \ldots, k$$

et les éléments de la matrice $H = \nabla^2 F(z_0)$ sont

$$h_{ij} = \frac{\partial^2 F}{\partial z_i \partial z_j}(z_0) = \sum_{l,p=1}^n \frac{\partial^2 f}{\partial x_l \partial x_p}(a)\, \frac{\partial \varphi_l}{\partial z_i}(z_0)\, \frac{\partial \varphi_p}{\partial z_j}(z_0) + \sum_{l=1}^n \frac{\partial f}{\partial x_l}(a)\, \frac{\partial^2 \varphi_l}{\partial z_i \partial z_j}(z_0).$$

Ce théorème donne un critère satisfaisant si la sous-variété $\mathcal{M}$ est donnée par une paramétrisation. Mais dans la plupart des applications, la sous-variété est donnée sous la forme $\mathcal{M} = \{x \in \mathbb{R}^n \mid g(x) = 0\}$. Le but est alors d'exprimer les conditions nécessaires et suffisantes du théorème 2.1 en termes de $g(x)$ et non de $\varphi(z)$.

La condition sur la première dérivée. La condition $\nabla f(a)^t \varphi'(z_0) = 0$, i.e. $\nabla f(a)^t \varphi'(z_0) v = 0$ pour tout $v \in \mathbb{R}^k$, signifie que le vecteur $\nabla f(a)$ est orthogonal à l'espace tangent $T_a\mathcal{M}$ (voir la formule (8.1) du paragraphe III.8). Il suffit d'utiliser la formule (8.2) du paragraphe III.8 qui dit que $T_a\mathcal{M} = \ker g'(a)$ et un théorème du cours d'Algèbre[1] pour conclure que $\nabla f(a) \in \operatorname{Im} g'(a)^t$. Ceci implique qu'il existe $\lambda = (\lambda_1, \ldots, \lambda_m)^t$ tel que

$$\nabla f(a) = g'(a)^t \lambda = \sum_{k=1}^{m} \lambda_k \nabla g_k(a). \tag{2.3}$$

Ici, on dénote par $g_k(x)$ les composantes de la fonction $g : \mathbb{R}^n \to \mathbb{R}^m$. Les paramètres $\lambda_1, \ldots, \lambda_m$ s'appellent les *multiplicateurs de Lagrange*. En introduisant la *fonction de Lagrange*

$$\mathcal{L}(x, \lambda) := f(x) - \lambda^t g(x) = f(x) - \sum_{k=1}^{m} \lambda_k g_k(x), \tag{2.4}$$

cette condition devient $\nabla_x \mathcal{L}(x, \lambda) = 0$, où ∇_x dénote le vecteur des dérivées partielles par rapport aux composantes de x. En résumé, un point $a \in \mathcal{M}$ satisfait la condition $\nabla f(a)^t \varphi'(z_0) = 0$ du théorème 2.1 si et seulement si (a, λ) est solution du système

$$\nabla_x \mathcal{L}(x, \lambda) = 0, \qquad g(x) = 0. \tag{2.5}$$

Il faut donc résoudre ce système à $n + m$ conditions pour $n + m$ inconnues ($x \in \mathbb{R}^n$ et $\lambda \in \mathbb{R}^m$).

La condition sur la deuxième dérivée. Essayons maintenant d'écrire la condition "H est définie (ou semi-définie) positive" du théorème 2.1 seulement à l'aide de la fonction $g(x)$.

Comme on a $g_k(\varphi(z)) = 0$, en dérivant par rapport à z_i, on obtient $\sum_l \frac{\partial g_k}{\partial x_l} \frac{\partial \varphi_l}{\partial z_i} = 0$ et en dérivant une deuxième fois, on obtient $\sum_{l,p} \frac{\partial^2 g_k}{\partial x_l \partial x_p} \frac{\partial \varphi_l}{\partial z_i} \frac{\partial \varphi_p}{\partial z_j} + \sum_l \frac{\partial g_k}{\partial x_l} \frac{\partial^2 \varphi_l}{\partial z_i \partial z_j} = 0$. En utilisant (2.3), l'élément (i, j) du deuxième terme de (2.2) est donc

$$\sum_l \frac{\partial f}{\partial x_l}(a)\, \frac{\partial^2 \varphi_l}{\partial z_i \partial z_j}(z_0) = \sum_{k,l} \lambda_k \frac{\partial g_k}{\partial x_l}(a)\, \frac{\partial^2 \varphi_l}{\partial z_i \partial z_j}(z_0) = -\sum_{k,l,p} \lambda_k \frac{\partial^2 g_k}{\partial x_l \partial x_p}(a)\, \frac{\partial \varphi_l}{\partial z_i}(z_0)\, \frac{\partial \varphi_p}{\partial z_j}(z_0).$$

L'expression à étudier devient alors

$$v^t H\, v = v^t \varphi'(z_0)^t \Big(\nabla^2 f(a) - \sum_{k=1}^{m} \lambda_k \nabla^2 g_k(a) \Big) \varphi'(z_0) v.$$

En posant $w = \varphi'(z_0) v$ et en utilisant $T_a\mathcal{M} = \operatorname{Im} \varphi'(z_0)$ (voir le paragraphe III.8), la condition "$v^t H\, v > 0$ pour $v \in \mathbb{R}^k$" devient

$$w^t \nabla_x^2 \mathcal{L}(a, \lambda)\, w = w^t \Big(\nabla^2 f(a) - \sum_{k=1}^{m} \lambda_k \nabla^2 g_k(a) \Big) w > 0 \quad \text{pour} \quad w \in T_a\mathcal{M}\ .$$

Celle-ci ne dépend plus de la paramétrisation $\varphi(z)$ mais uniquement de $g(x)$.

[1] *Rappel du cours d'Algèbre.* Pour une matrice arbitraire A on a $(\ker A)^\perp = \operatorname{Im} A^t$. On utilise ici la notation $V^\perp = \{w \in \mathbb{R}^n \mid w^t v = 0 \text{ pour tout } v \in V\}$.
Démonstration. Ceci suit des équivalences $v \in \ker A \;\Leftrightarrow\; Av = 0 \;\Leftrightarrow\; w^t A v = 0$ pour tout $w \;\Leftrightarrow\; v \perp A^t w$ pour tout $w \;\Leftrightarrow\; v \perp \operatorname{Im} A^t \;\Leftrightarrow\; v \in (\operatorname{Im} A^t)^\perp$ et du fait que $(V^\perp)^\perp = V$.

Nous avons alors démontré le résultat suivant:

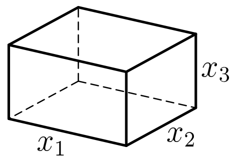
Théorème 2.2 *Soient $f : \mathbb{R}^n \to \mathbb{R}$ suffisamment différentiable, $\mathcal{M} = \{x \in \mathbb{R}^n \mid g(x) = 0\}$ une sous-variété, $a \in \mathcal{M}$ et $\lambda \in \mathbb{R}^m$. Avec $\mathcal{L}(x, \lambda)$ donnée par (2.4), on a alors*

$$\left\{ \begin{array}{c} \nabla_x \mathcal{L}(a,\lambda) = 0 \textit{ et} \\ w^t \nabla_x^2 \mathcal{L}(a,\lambda)\, w > 0 \\ \textit{pour } w \in T_a\mathcal{M},\ w \neq 0 \end{array} \right\} \Rightarrow \left\{ \begin{array}{c} f|_{\mathcal{M}} \textit{ possède un} \\ \textit{minimum relatif en } a \end{array} \right\} \Rightarrow \left\{ \begin{array}{c} \nabla_x \mathcal{L}(a,\lambda) = 0 \textit{ et} \\ w^t \nabla_x^2 \mathcal{L}(a,\lambda)\, w \geq 0 \\ \textit{pour } w \in T_a\mathcal{M}. \end{array} \right\} \quad \square$$

Corollaire 2.3 *Situation comme dans le théorème 2.2. Alors on a*

$$\left\{ \begin{array}{c} \nabla_x \mathcal{L}(a,\lambda) = 0 \textit{ et} \\ w^t \nabla_x^2 \mathcal{L}(a,\lambda)\, w < 0 \\ \textit{pour } w \in T_a\mathcal{M},\ w \neq 0 \end{array} \right\} \Rightarrow \left\{ \begin{array}{c} f|_{\mathcal{M}} \textit{ possède un} \\ \textit{maximun relatif en } a \end{array} \right\} \Rightarrow \left\{ \begin{array}{c} \nabla_x \mathcal{L}(a,\lambda) = 0 \textit{ et} \\ w^t \nabla_x^2 \mathcal{L}(a,\lambda)\, w \leq 0 \\ \textit{pour } w \in T_a\mathcal{M}. \end{array} \right\} \quad \square$$

Exemple 2.4 Considérons un parallélépipède droit dont la largeur, la profondeur et la hauteur sont respectivement x_1, x_2 et x_3. Le problème consiste à maximiser le volume $f(x_1, x_2, x_3) = x_1x_2x_3$ tout en gardant le périmètre constant et égal à P (i.e. la sous-variété $\mathcal{M}$ est donnée par $g(x_1, x_2, x_3) = 4x_1 + 4x_2 + 4x_3 - P = 0$). La fonction de Lagrange pour ce problème est



$$\mathcal{L}(x, \lambda) = x_1x_2x_3 - \lambda(4x_1 + 4x_2 + 4x_3 - P)$$

et la condition nécessaire (2.5) devient

$$\begin{array}{l} x_2x_3 - 4\lambda = 0 \\ x_1x_3 - 4\lambda = 0 \\ x_1x_2 - 4\lambda = 0 \end{array} \qquad 4x_1 + 4x_2 + 4x_3 - P = 0.$$

Les solutions de ce système sont $(P/4, 0, 0)$, $(0, P/4, 0)$, $(0, 0, P/4)$ avec $\lambda = 0$ dans chaque cas et $(P/12, P/12, P/12)$ avec $\lambda = P^2/576$. Il est facile de deviner que le volume est maximal pour $x_1 = x_2 = x_3 = P/12$, mais vérifions tout de même la condition suffisante du corollaire 2.3. Pour ceci, nous calculons

$$\nabla_x^2 \mathcal{L}(x, \lambda) = \begin{pmatrix} 0 & x_3 & x_2 \\ x_3 & 0 & x_1 \\ x_2 & x_1 & 0 \end{pmatrix}$$

ainsi que l'espace tangent pour $\mathcal{M} = \{(x_1, x_2, x_3)^t \mid g(x_1, x_2, x_3) = 0\}$:

$$T_a\mathcal{M} = \{(v_1, v_2, v_3)^t \mid v_1 + v_2 + v_3 = 0\} = \langle (1, -1, 0)^t,\ (1, 0, -1)^t \rangle.$$

La condition $w^t \nabla_x^2 \mathcal{L}(a, \lambda)\, w < 0$ pour $w \in T_a\mathcal{M},\ w \neq 0$ est alors équivalente à la propriété que

$$\begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} 0 & x_3 & x_2 \\ x_3 & 0 & x_1 \\ x_2 & x_1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} -2x_3 & x_1 - x_2 - x_3 \\ x_1 - x_2 - x_3 & -2x_2 \end{pmatrix}$$

soit définie négative. Cette condition est satisfaite pour le point $(P/12, P/12, P/12)$. Pour les autres trois points critiques, les deux valeurs propres de cette matrice sont de signes opposés.

Exemple 2.5 Cherchons à minimiser la fonction $f(x_1, x_2, x_3) = x_1 - x_2 - x_3$ sur la sous-variété $\mathcal{M} = \{(x_1, x_2, x_3)^t \,|\, x_1^2 + 2x_2^2 - 1 = 0,\ 3x_1 - 4x_3 = 0\}$. Cette dernière représente l'intersection d'un cylindre vertical avec un plan, c.-à-d. une ellipse dans l'espace. La fonction de Lagrange est

$$\mathcal{L}(x_1, x_2, x_3, \lambda_1, \lambda_2) = x_1 - x_2 - x_3 - \lambda_1(x_1^2 + 2x_2^2 - 1) - \lambda_2(3x_1 - 4x_3)$$

et la condition nécessaire (2.5) devient

$$\begin{aligned} 1 - 2\lambda_1 x_1 - 3\lambda_2 &= 0 \\ -1 - 4\lambda_1 x_2 &= 0 \\ -1 + 4\lambda_2 &= 0 \end{aligned} \qquad \begin{aligned} x_1^2 + 2x_2^2 - 1 &= 0 \\ 3x_1 - 4x_3 &= 0 \end{aligned}$$

dont les solutions sont

$$\lambda_2 = \frac{1}{4}, \quad \lambda_1 = \pm\frac{3}{8}, \quad x_1 = \pm\frac{1}{3}, \quad x_2 = \mp\frac{2}{3}, \quad x_3 = \pm\frac{1}{4}. \tag{2.6}$$

On aimerait déterminer laquelle de ces deux solutions représente un minimum (ou maximum) de la fonction f. Nous calculons alors

$$\nabla_x^2 \mathcal{L}(x, \lambda) = -\lambda_1 \begin{pmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

et l'espace tangent pour $\mathcal{M}$ (au points de la solution (2.6))

$$T_a\mathcal{M} = \{(v_1, v_2, v_3)^t \,|\, 2x_1v_1 + 4x_2v_2 = 0,\ 3v_1 - 4v_3 = 0\} = \langle (4, 1, 3)^t \rangle.$$

Avec $w := (4, 1, 3)^t$ on a $w^t \nabla_x^2 \mathcal{L}(x, \lambda)\, w = -36\lambda_1$. La fonction f est donc minimale sur $\mathcal{M}$ au point $(-1/3, 2/3, -1/4)^t$ où λ_1 est négatif et elle est maximale sur $\mathcal{M}$ au point $(1/3, -2/3, 1/4)^t$.

IV.3 Contraintes en forme des équations et inéquations

On cherche des conditions nécessaires et suffisantes pour un problème encore plus général qui est le suivant: soient $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$ et $h : \mathbb{R}^n \to \mathbb{R}^k$ continûment différentiables; trouver les minima relatifs du problème

$$f(x) \to \min \qquad \text{sur} \qquad K = \{x \in \mathbb{R}^n \mid g(x) = 0,\ h(x) \geq 0\}. \tag{3.1}$$

La condition "$h(x) \geq 0$" signifie que tous les composantes de $h(x)$ doivent être non négatives, c.-à-d. $h_j(x) \geq 0$ pour $j = 1, \ldots, k$. Supposons pour le moment que K soit uniquement donné par des inéquations $h(x) \geq 0$.

Exemple 3.1 Considérons la fonction

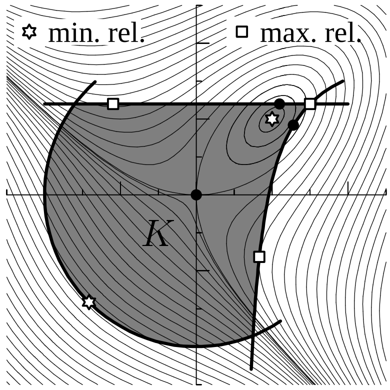
$$f(x_1, x_2) = x_1^3 + x_2^3 - 3x_1x_2$$

qui nous est familière de l'exemple 1.3 et cherchons ses minima (et maxima) relatifs sur l'ensemble

$$K = \{(x_1, x_2) \mid 4 - x_1^2 - x_2^2 \geq 0,\ \ -x_2 + 1.2 \geq 0,\ (x_1 - 0.5)(x_2 - 2.2) + 1 \geq 0\}.$$

Pour résoudre ce problème, nous le séparons en plusieurs sous-problèmes.

- Étudions d'abord les points critiques de la fonction $f(x_1, x_2)$ sans tenir compte de l'ensemble K. Ceux qui sont par hasard dans K, sont aussi des extrema de la fonction $f|_K$ (restriction de f sur K). Dans notre situation (voir exemple 1.3), ceci est le cas pour le minimum relatif $a = (1, 1)$ ainsi que pour le point de selle $a = (0, 0)$.



- Ensuite, il faut considérer le bord de K. Comme il consiste en plusieurs courbes lisses, commençons à étudier f sur le cercle $\mathcal{M}_1 = \{(x_1, x_2) \mid 4 - x_1^2 - x_2^2 = 0\}$. À l'aide de la fonction de Lagrange

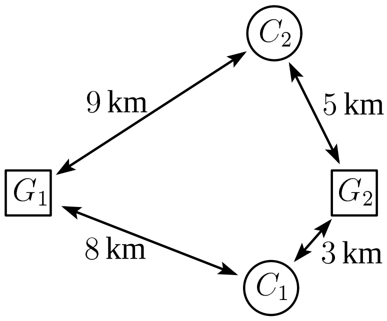
  $$\mathcal{L}(x_1, x_2, \lambda) = x_1^3 + x_2^3 - 3x_1x_2 - \lambda(4 - x_1^2 - x_2^2),$$

  nous trouvons $(-\sqrt{2}, -\sqrt{2})$ et trois autres solutions à l'extérieur de K. Le dessin nous montre qu'il s'agit d'un minimum (global). En procédant de la même manière avec les autres courbes du bord de K, nous obtenons $(-\sqrt{1.2}, 1.2)$ et $(0.8315, -0.8159)$ comme maxima relatifs, et les points $(\sqrt{1.2}, 1.2)$ et $(1.2797, 0.9175)$ qui sont des minima sur ∂K mais pas sur K.
- Finalement, nous devrons étudier les sommets de K. En inspectant la figure, nous voyons que le point $(1.5, 1.2)$ est un maximum relatif de f sur K.

IV.4 Programmation linéaire

Les problèmes convexes les plus simples et les plus importants sont ceux où toutes les fonctions sont linéaires (affines). Nous commençons par un exemple de petite dimension qui permet une interprétation et une résolution graphique. L'intérêt principal des paragraphes IV.4 et IV.5 est de développer un algorithme qui permet de résoudre n'importe quel problème de la programmation linéaire dans une dimension arbitraire.

Exemple 4.1 Une entreprise possède 2 garages G_1 et G_2 avec 8 et 6 camions respectivement. Elle travaille sur deux chantiers C_1 et C_2. Le chantier C_1 exige 4 camions et le chantier C_2 exige 7 camions. Les distances sont données sur le schéma d'à côté. Comment faut-il organiser les trajets pour minimiser les consommations d'essence?



Pour une formulation mathématique du problème, notons par x_1, x_2, x_3 et x_4 le nombre de camions allant de G_1 à C_1, de G_1 à C_2, G_2 à C_1 et de G_2 à C_2 respectivement.

Nous avons les conditions

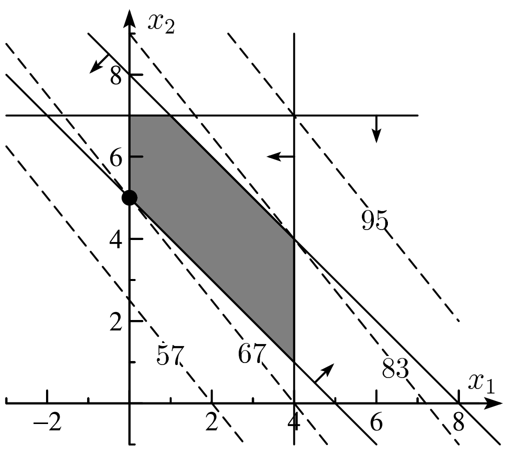
$$x_1 + x_2 \leq 8\ , \quad x_3 + x_4 \leq 6\ , \quad x_1 + x_3 = 4\ , \quad x_2 + x_4 = 7 \quad \text{et} \quad x_i \geq 0$$

et la fonction objective est

$$8x_1 + 9x_2 + 3x_3 + 5x_4 \to \min .$$

On peut simplifier le problème en éliminant x_3 et x_4 car $x_3 = 4 - x_1$ et $x_4 = 7 - x_2$. On obtient alors

$$\begin{aligned} 5x_1 + 4x_2 + 47 &\to \min \\ x_1 + x_2 &\leq 8 \\ -x_1 - x_2 &\leq -5 \\ x_1 \leq 4, &\quad x_1 \geq 0 \\ x_2 \leq 7, &\quad x_2 \geq 0. \end{aligned}$$



En regardant les droites où $f(x) = 5x_1 + 4x_2 + 47$ est constante (lignes traitillées), on trouve immédiatement la solution du problème: $x_1 = 0$, $x_2 = 5$, $x_3 = 4$, $x_4 = 2$.

Formulation du problème général. Nous considérons la résolution du problème suivant:

$$\begin{array}{l} c_1x_1 + c_2x_2 + \ldots + c_nx_n \to \max \\ a_{11}x_1 + \ldots + a_{1n}x_n \le b_1 \\ \qquad \ldots \\ a_{m1}x_1 + \ldots + a_{mn}x_n \le b_m \\ x_1 \ge 0, \ \ldots, \ x_n \ge 0 \end{array} \qquad \text{ou} \qquad \begin{array}{l} c^tx \to \max \\ Ax \le b \\ x \ge 0 \end{array} \tag{4.1}$$

D'autres problèmes de la programmation linéaire peuvent être écrits sous cette forme:

- La condition $c^tx \to \min$ est équivalente à $(-c)^tx \to \max$;
- Une équation peut être remplacée par deux inéquations; ou mieux encore: éliminer des variables comme dans l'exemple précédent;
- Une variable x_1, sans restriction sur son signe, peut être écrite comme $x_1 = x_1^+ - x_1^-$ où $x_1^+ \ge 0$ et $x_1^- \ge 0$.

Un algorithme possible mais pas recommandé. L'ensemble, où on cherche à maximiser la fonction objective $f(x) = c^tx$, est le polyèdre $K = \{x \in \mathbb{R}^n \mid Ax \le b\,,\, x \ge 0\}$. Comme $f(x)$ est linéaire, une solution est certainement sur un des sommets de K. Pour la trouver, on pourrait procéder de la manière suivante:

- calculer tous les sommets du polyèdre (le nombre de sommets est fini);
- déterminer le sommet où $f(x) = c^tx$ est maximal.

L'algorithme que nous allons présenter dans le paragraphe suivant est une procédure intelligente qui permet d'éviter le calul de *tous* les sommets du polyèdre. Néanmoins, nous sommes obligés de formuler mathématiquement les sommets du polyèdre. C'est notre but pour le reste de ce paragraphe.

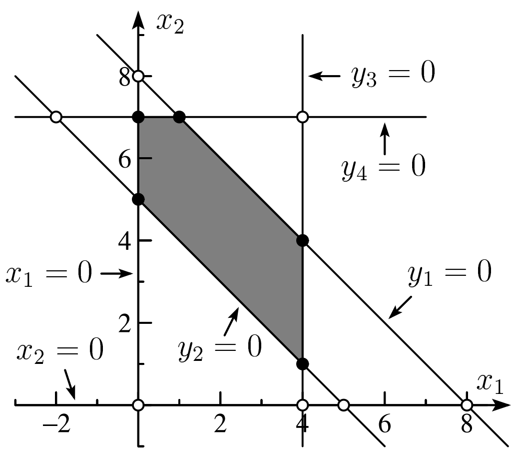
Variables d'écart (variables artificielles). On pose $y = b - Ax$ où $y \in \mathbb{R}^m$. Pour un $x \in \mathbb{R}^n$ donné, y_i mesure son *écart* (distance) à l'hyperplan $\sum_{j=1}^n a_{ij}\xi_j - b_i = 0$. Le problème (4.1) est donc équivalent à

$$c^tx \to \max, \quad Ax + y = b, \quad x \ge 0, \quad y \ge 0. \tag{4.2}$$

Cette reformulation nous permet de calculer systématiquement les sommets du polyèdre $K = \{x \in \mathbb{R}^n \mid Ax \le b\,,\, x \ge 0\}$. Illustrons ceci avec le problème de l'exemple 4.1.

Exemple 4.2 Nous introduisons des variables d'écart $y_1, \ldots, y_4$ pour les quatre inéquations du problème de l'exemple 4.1. Ceci donne

$$\begin{aligned} x_1 + x_2 + y_1 &= 8 \\ -x_1 - x_2 + y_2 &= -5 \\ x_1 + y_3 &= 4 \\ x_2 + y_4 &= 7 \\ x_i \ge 0, \quad y_i &\ge 0 \end{aligned}$$



et on a 4 équations à 6 inconnues. Chacune des conditions $x_i = 0$ et $y_i = 0$ correspond à une droite sur le dessin d'à côté. Si l'on pose deux

variables parmi $x_1,\ x_2,\ y_1, \ldots, \ y_4$ égal à zéro, on obtient les intersections de deux droites et, parmi elles, les sommets du polyèdre. Il y a $\binom{6}{2} = 15$ possibilités d'annuler deux variables. On obtient ainsi les 5 sommets du polyèdre et les 7 autres intersections; 3 couples ne donnent pas de solution (droites parallèles).

Cet exemple nous montre que, pour le calcul des sommets du polyèdre, on traîte les variables x_i et y_i équitablement. Il n'y a donc pas de raison de distinguer les deux types de variables. Rassemblons-les donc toutes dans un vecteur $z = (x_1, \ldots, \ x_n,\ y_1, \ldots, \ y_m)^t$ et écrivons le problème (4.2) sous la forme

$$(c^t,\ 0^t)\, z \to \max, \quad (A,\ I)\, z = b, \quad z \ge 0 \quad \text{avec} \quad (A,\ I) = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ -1 & -1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{4.3}$$

(pour le système de l'exemple 4.2). Nous constatons que l'annulation de deux variables parmi $x_1,\ x_2,\ y_1, \ldots, \ y_4$ correspond à une intersection réelle si et seulement si la matrice carrée, obtenue en supprimant les deux colonnes correspondantes, est inversible. Si, en plus, les composantes de la solution de $(A,\ I)\, z = b$ sont non négatives, alors cette intersection est un sommet du polyèdre.

Dans le problème (4.3) nous écrivons de nouveau c^t au lieu de $(c^t,\ 0^t)$, A au lieu de $(A,\ I)$ et x au lieu de z. Le problème devient alors

$$\begin{aligned} & c^t x \to \max \\ & Ax = b \\ & x \ge 0 \end{aligned} \tag{4.4}$$

où $x \in \mathbb{R}^n$, $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$ et $n \ge m$. Nous supposons (sans perdre la généralité) que le rang de la matrice A soit m (maximal). Sinon, soit le système $Ax = b$ ne possède pas de solution soit une ou plusieurs équations du système peuvent être supprimées sans changer les solutions.

Notations. Par la suite, nous noterons $A_j \in \mathbb{R}^m$ la $j$ème colonne de la matrice A. Nous partitionnons les indices $\{1, \ldots, n\} = B \cup R$ en deux sous-ensembles disjoints dont $B = \{i_1, \ldots, i_m\}$ contient m éléments et $R = \{j_1, \ldots, j_{n-m}\}$ contient le reste. Nous partitionnons également les vecteurs c, x et la matrice A de la manière suivante:

$$\begin{aligned} A_B &= (A_{i_1}, \ldots, \ A_{i_m}) & \qquad A_R &= (A_{j_1}, \ldots, \ A_{j_{n-m}}) \\ x_B &= (x_{i_1}, \ldots, \ x_{i_m})^t & \qquad x_R &= (x_{j_1}, \ldots, \ x_{j_{n-m}})^t \\ c_B &= (c_{i_1}, \ldots, \ c_{i_m})^t & \qquad c_R &= (c_{j_1}, \ldots, \ c_{j_{n-m}})^t. \end{aligned}$$

Ainsi, A_B est une matrice carrée de dimension m. Le problème (4.4) devient alors

$$\begin{aligned} & c_B^t x_B + c_R^t x_R \to \max \\ & A_B x_B + A_R x_R = b \\ & x_B \ge 0, \ \ x_R \ge 0. \end{aligned} \tag{4.5}$$

Définition 4.3 (base) Un sous-ensemble $B = \{i_1, \ldots, i_m\} \subset \{1, \ldots, n\}$ à m éléments s'appelle *une base*, si A_B est inversible, c.-à-d. si $\det A_B \neq 0$. Pour une base B, le vecteur

$$x = \begin{pmatrix} x_B \\ x_R \end{pmatrix} = \begin{pmatrix} A_B^{-1} b \\ 0 \end{pmatrix} \tag{4.6}$$

s'appelle *solution de base* (associée à B). Elle est dite *admissible* si $A_B^{-1} b \ge 0$.

L'interprétation géométrique de cette définition est la suivante: l'ensemble $\{x \mid Ax = b\}$ représente un espace affine de dimension $n - m$ (car A est une matrice $m \times n$ de rang maximal m) et l'ensemble $\{x \mid Ax = b,\ x \geq 0\}$ est un polyèdre dans cet espace. Pour une base B du problème, la condition $x_R = (x_{j_1}, \ldots, x_{j_{n-m}})^t = 0$ représente l'intersection de $n - m$ hyperplans qui sont des faces du polyèdre. Les solutions de base admissibles sont donc en bijection avec les sommets de ce polyèdre.

Par exemple, $\{y_1, y_2, y_3, y_4\}$ (ou plus précisément l'ensemble $\{3, 4, 5, 6\}$) est une base du problème de l'exemple 4.2. Elle correspond à l'origine qui est l'intersection de $x_1 = 0$ et de $x_2 = 0$. La base $\{x_1, x_2, y_2, y_3\}$ correspond au sommet $(4, 4)$ qui est l'intersection de $y_1 = 0$ et de $y_3 = 0$. L'ensemble $\{x_1, x_2, y_3, y_4\}$ ne constitue pas une base parce que la matrice formée par les deux premières et les deux dernières colonnes de (4.3) n'est pas inversible (les droites $y_1 = 0$ et $y_2 = 0$ n'ont pas d'intersection).

Reformulation du problème. Par élimination des variables x_B, c.-à-d. en utilisant la relation $x_B = -A_B^{-1} A_R x_R + A_B^{-1} b$, le problème (4.5) devient équivalent à

$$\begin{array}{c} (-c_B^t A_B^{-1} A_R + c_R^t) x_R + c_B^t A_B^{-1} b \to \max \\ A_B^{-1} A_R x_R \leq A_B^{-1} b \\ x_R \geq 0 \end{array} \qquad \text{ou} \qquad \begin{array}{c} u^t x_R - z \to \max \\ S x_R \leq t \\ x_R \geq 0 \end{array} \tag{4.7}$$

où $S = A_B^{-1} A_R$, $t = A_B^{-1} b$, $u^t = -c_B^t A_B^{-1} A_R + c_R^t$ et $z = -c_B^t A_B^{-1} b$.

Tableau du simplexe. Nous rassemblons toute l'information de la formulation (4.7) dans le tableau suivant:

	R	
B	S	t
	u^t	z

Voici quelques tableaux du simplexe pour le problème de l'exemple 4.2:

	x_1	x_2	
y_1	1	1	8
y_2	−1	−1	−5
y_3	1	0	4
y_4	0	1	7
	−5	−4	0

	y_1	y_3	
x_1	0	1	4
x_2	1	−1	4
y_2	1	0	3
y_4	−1	1	3
	4	1	36

	x_1	y_2	
x_2	1	−1	5
y_1	0	1	3
y_3	1	0	4
y_4	−1	1	2
	−1	−4	20

(4.8)

La solution de base pour le premier tableau n'est pas admissible (le vecteur $t = A_B^{-1} b$ possède des composantes négatives) mais celles du deuxième et du troisième le sont et correspondent à des sommets du polyèdre (voir le dessin de l'exemple 4.2).

Théorème 4.4 (critère du simplexe) *Soit B une base du problème*

$$\begin{array}{l} c^t x \to \max \\ Ax = b \\ x \geq 0 \end{array}$$

et

	R	
B	S	t
	u^t	z

le tableau du simplexe correspondant.

Si $t \geq 0$ et $u \leq 0$, alors $x_B = t$, $x_R = 0$ est solution du problème.

Démonstration. Comme $t \geq 0$, la solution de base associée à B est admissible. De plus, pour un point x satisfaisant $Ax = b$ et $x \geq 0$, la conditon $u \leq 0$ implique

$$f(x) = \sum_{j \in R} u_j x_j - z \leq -z = f\left(\begin{pmatrix} t \\ 0 \end{pmatrix}\right).$$

Cette solution de base est alors une solution du problème d'optimisation. □

Le troisième tableau de (4.8) vérifie le critère du simplexe ($t \geq 0$ et $u \leq 0$). La solution de base associée $x_1 = 0$ et $x_2 = 5$ est donc une solution optimale du problème de l'exemple 4.2.

Pour se convaincre que l'algorithme (non recommandé) du début de ce paragraphe n'est pas réalisable, considérons le cas où $n = 20$ et $m = 10$ (en pratique on est confronté à des dimensions beaucoup plus grande). Il y a $\binom{20}{10} = 184\,756$ possibilités de choisir 10 parmi les 20 colonnes de la matrice A. Le calcul de toutes les solutions de base pourrait ainsi exiger la résolution d'environ $180\,000$ systèmes linéaires de 10 équations à 10 inconnues.

IV.5 L'algorithme du simplexe

> The author wishes to acknowledge that his work on this subject stemmed from discussions in the spring of 1947 with Marshall K. Wood, in connection with Air Force programming methods. The general nature of the "simplex" approach (...) was stimulated by discussions with Leonid Hurwicz. (G.B. Dantzig 1951)

> L'idée de suivre les arêtes d'un polyèdre pour optimiser une fonction linéaire sous contraintes linéaires a déjà été envisagée par J. Fourier[2]. Dantzig cependant la systématise.
> (J.F. Maurras 2002[3])

L'algorithme du simplexe a été publié par Dantzig comme chapitre XXI dans un compte rendu d'une conférence[4]. C'est une procédure itérative permettant d'effectuer une exploration dirigée des sommets du polyèdre en suivant des arêtes jusqu'au sommet qui maximise la fonction objective. Cet algorithme est parmi les "Top 10 Algorithms of the 20th Century" (SIAM News, Vol. 33, Nr. 4, May 2000).

Le théorème suivant donne les formules qui, en partant d'un tableau du simplexe, nous permet d'obtenir le tableau associé à un sommet voisin.

Théorème 5.1 *Soit B une base du problème (4.4), $R = \{1, \ldots, n\} \setminus B$, $k \in B$, $l \in R$ et*

$$\begin{array}{|c|c|c|} \hline & R & \\ \hline B & S & t \\ \hline & u^t & z \\ \hline \end{array} \quad = \quad \begin{array}{|c|cc|c|} \hline & l & j & \\ \hline k & \boxed{s_{kl}} & s_{kj} & t_k \\ i & s_{il} & s_{ij} & t_i \\ \hline & u_l & u_j & z \\ \hline \end{array} \tag{5.1}$$

le tableau du simplexe correspondant. Si $s_{kl} \neq 0$ (pivot), alors

- $\widehat{B} = B \cup \{l\} \setminus \{k\}$ *est une base.*

[2]J.B.J. Fourier, *Solution d'une question particulière du calcul des inégalités*, Nouveau Bulletin des Sciences par la Société Philomathique de Paris (1826) 99–100.

[3]J.F. Maurras, *Programmation Linéaire, Complexité, Séparation et Optimisation*, Mathématiques & Applications 38, Springer, Paris, 2002.

[4]G.B. Dantzig, *Maximization of a linear function of variables subject to linear inequalities*, Activity Analysis of Production and Allocation (T.C. Koopmans, ed.), Wiley, New York, 1951, 339–347.

- *Le tableau qui correspond à $\widehat{B}$ est donné par*

$$
\begin{array}{|c|c|c|}
\hline
 & \widehat{R} & \\
\hline
\widehat{B} & \widehat{S} & \widehat{t} \\
\hline
 & \widehat{u}^t & \widehat{z} \\
\hline
\end{array}
\quad = \quad
\begin{array}{|c|cc|c|}
\hline
 & k & j & \\
\hline
l & \frac{1}{s_{kl}} & \frac{s_{kj}}{s_{kl}} & \frac{t_k}{s_{kl}} \\
i & -\frac{s_{il}}{s_{kl}} & s_{ij} - \frac{s_{kj}s_{il}}{s_{kl}} & t_i - \frac{t_k s_{il}}{s_{kl}} \\
\hline
 & -\frac{u_l}{s_{kl}} & u_j - \frac{s_{kj}u_l}{s_{kl}} & z - \frac{t_k u_l}{s_{kl}} \\
\hline
\end{array}
$$

Voici une astuce pour retenir cette formule:

$$
\begin{pmatrix} p & q \\ r & s \end{pmatrix} \longrightarrow \begin{pmatrix} \widehat{p} & \widehat{q} \\ \widehat{r} & \widehat{s} \end{pmatrix} = \begin{pmatrix} \frac{1}{p} & \frac{q}{p} \\ -\frac{r}{p} & s - \frac{rq}{p} \end{pmatrix}.
$$

Démonstration. La $k$ème composante de l'équation $x_B + Sx_R = t$ est $x_k + \sum_{j \in R} s_{kj}x_j = t_k$. Comme $s_{kl} \neq 0$, on peut en extraire x_l ($l \in R$, fixé) et on obtient

$$
x_l + \frac{1}{s_{kl}} x_k + \sum_{j \in R \setminus \{l\}} \frac{s_{kj}}{s_{kl}} x_j = \frac{t_k}{s_{kl}}
$$

(voir la ligne "l" du tableau pour $\widehat{B}$). En remplaçant x_l de cette formule dans la $i$ème composante de $x_B + Sx_R = t$ ($i \in B \setminus \{k\}$), on obtient

$$
x_i + s_{il}\Big(-\frac{1}{s_{kl}} x_k - \sum_{j \in R \setminus \{l\}} \frac{s_{kj}}{s_{kl}} x_j + \frac{t_k}{s_{kl}}\Big) + \sum_{j \in R \setminus \{l\}} s_{ij}x_j = t_i \quad \text{ou} \quad x_i + \widehat{s}_{ik}x_k + \sum_{j \in R \setminus \{l\}} \widehat{s}_{ij}x_j = \widehat{t}_i
$$

ce qui vérifie la ligne "i" du tableau pour $\widehat{B}$. La dernière ligne du tableau est obtenue de la même manière. □

Les formules du théorème précédent donnent un moyen simple et efficace pour calculer des tableaux du simplexe à partir d'un tableau connu. Il reste à éviter le calcul des tableaux inutiles.

Algorithme du simplexe. Soit donné un tableau (5.1) avec $t \geq 0$, c.-à-d. la solution de base associée est admissible. Alors, le programme suivant nous permet de trouver une solution optimale:

begin
if $(u \leq 0)$ **then**
  on a trouvé une solution du problème, **stop**
else
  choisir $l \in R$ avec $u_l > 0$
  if ($s_{il} \leq 0$ pour tout $i \in B$) **then**
    pas de solution, la fonction objective n'est pas majorée, **stop** (voir explication 1)
  else
    choisir $k \in B$ tel que $\frac{t_k}{s_{kl}} = \min\{\frac{t_i}{s_{il}} \mid i \in B, s_{il} > 0\}$;
    échanger k et l avec les formules du théorème précédent;
    on obtient un nouveau tableau avec base $\widehat{B} = B \cup \{l\} \setminus \{k\}$ qui satisfait
    $\widehat{t} \geq 0$ et $-\widehat{z} \geq -z$ (voir explication 2) **go to begin**
  end if
end if
end

Explications.

1. *Si $u_l > 0$ (pour un $l \in R$) et $s_{il} \le 0$ pour tout $i \in B$ alors la fonction objective n'est pas majorée sur l'ensemble des points admissibles.*

Démonstration. Considérons le rayon $x_R(\alpha)$ défini par $x_l(\alpha) = \alpha$ ($\alpha \ge 0$) et $x_j(\alpha) = 0$ pour $j \in R \setminus \{l\}$. Alors $x_R(\alpha) \ge 0$ et $Sx_R(\alpha) = \alpha\,(\ldots, s_{il}, \ldots)^t \le 0 \le t$. Les points du rayon $x_R(\alpha)$ sont donc admissible pour tout $\alpha \ge 0$ et

$$f(x(\alpha)) = u^t x_R(\alpha) - z = u_l\alpha - z \to \infty \quad \text{si} \quad \alpha \to \infty. \qquad \square$$

2. *Le choix de l et de k dans l'algorithme conduit à un tableau avec $\widehat{t} \ge 0$ et avec $-\widehat{z} \ge -z$.*

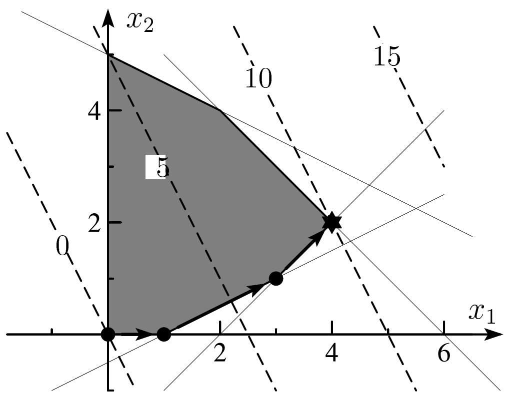
Démonstration. Nous avons $\widehat{t_l} = \frac{t_k}{s_{kl}} \ge 0$ parce que $t_k \ge 0$ et $s_{kl} > 0$ et pour $i \in B \setminus \{k\}$

$$\widehat{t_i} = t_i - \frac{t_k s_{il}}{s_{kl}} \quad \begin{cases} \ge t_i \ge 0 & \text{si} \quad s_{il} \le 0 \\ = s_{il}\left(\frac{t_i}{s_{il}} - \frac{t_k}{s_{kl}}\right) \ge 0 & \text{si} \quad s_{il} > 0. \end{cases}$$

De plus, on a $-\widehat{z} = -z + \frac{t_k u_l}{s_{kl}} \ge -z$ car $u_l > 0$, $t_k \ge 0$ et $s_{kl} > 0$. $\square$

Exemple 5.2 Considérons le problème

$$\begin{aligned} &2x_1 + x_2 \to \max \\ &x_1 + 2x_2 \le 10 \\ &x_1 + x_2 \le 6 \\ &x_1 - x_2 \le 2 \\ &x_1 - 2x_2 \le 1 \\ &x_1 \ge 0, \;\; x_2 \ge 0 \end{aligned}$$



et appliquons l'algorithme du simplexe.

Nous introduisons les quatre variables d'écart y_1, y_2, y_3, y_4 pour les quatre inéquations. Ceci donne un problème de programmation linéaire avec $m = 4$ équations et $n = 6$ variables non négatives. Les variables y_1, y_2, y_3, y_4 constituent une base, dont la solution associée $x_1 = x_2 = 0$ est admissible. Cette base nous donne un tableau de départ pour l'algorithme du simplexe:

	x_1	x_2	
y_1	1	2	10
y_2	1	1	6
y_3	1	-1	2
y_4	$\boxed{1}$	-2	1
	2	1	0

	y_4	x_2	
y_1	-1	4	9
y_2	-1	3	5
y_3	-1	$\boxed{1}$	1
x_1	1	-2	1
	-2	5	-2

	y_4	y_3	
y_1	3	-4	5
y_2	$\boxed{2}$	-3	2
x_2	-1	1	1
x_1	-1	2	3
	3	-5	-7

	y_2	y_3	
y_1	$-3/2$		2
y_4	$1/2$	$-3/2$	1
x_2	$1/2$		2
x_1	$1/2$		4
	$-3/2$	$-1/2$	-10

Dans le premier tableau, le deux composantes de u sont positives et on a le choix entre x_1 et x_2, c.-à-d. on peut soit aller vers le sommet $(1, 0)$ (c'était notre choix) soit vers le sommet $(0, 5)$. Dans les autres étapes, le pivot est unique. Il nous faut trois itérations pour un tableau qui satisfait le critère du simplexe. La solution du problème est $x_1 = 4$, $x_2 = 2$ et la valeur maximale de la fonction objective est $f(x_{\text{opt}}) = 10$.

Construction d'une solution de base admissible. L'algorithme du simplexe nécessite la connaissance d'une base avec $t \geq 0$ (solution de base admissible). Dans un problème

$$\begin{array}{lcl} c^t x \to \max & & \sum_{i=1}^n c_i x_i \to \max \\ Ax \leq b & \text{ou de manière équivalente} & y_i + \sum_{j=1}^n a_{ij} x_j = b_i \\ x \geq 0 & & x_i \geq 0, \;\; y_i \geq 0 \end{array} \tag{5.2}$$

avec $b \geq 0$, les variables $y_1, \ldots, y_m$ forment une base dont la solution associée est admissible. On peut alors immédiatement appliquer l'algorithme du simplexe.

Si, par contre, au moins une composante du vecteur b est négative, il faut une astuce supplémentaire pour trouver un tableau de départ. L'idée est de supprimer les inéquations avec $b_i < 0$ (pour que l'origine devienne un point admissible, voir le dessin de l'exemple 5.3) et ensuite de minimiser l'écart de celles-ci. Cela revient à considérer le *problème auxiliaire*

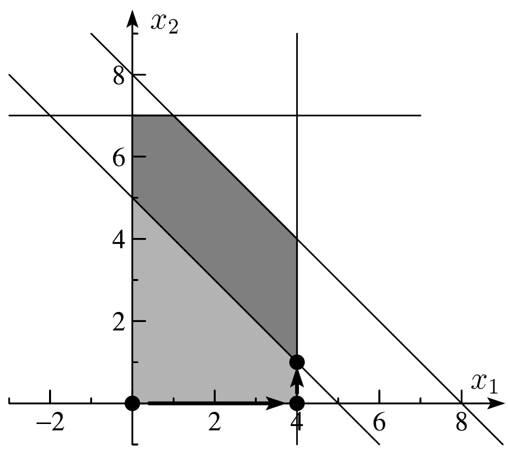
$$\begin{array}{ll} y_i + \sum_{j=1}^n a_{ij} x_j = b_i & \text{si} \quad b_i \geq 0 \\ y_i - z_i + \sum_{j=1}^n a_{ij} x_j = b_i & \text{si} \quad b_i < 0 \\ y_i \geq 0, \;\; z_i \geq 0, \;\; x_i \geq 0 & \\ -\sum_i z_i \to \max & \end{array} \tag{5.3}$$

Ce problème auxiliaire a les propriétés suivantes:

- L'ensemble des variables $\{y_i \mid b_i \geq 0\} \cup \{z_i \mid b_i < 0\}$ forme une base de (5.3) avec solution associée $x_i = 0$, $y_i = 0$ si $b_i < 0$, $y_i = b_i$ si $b_i \geq 0$ et $z_i = -b_i$ si $b_i < 0$. Cette solution satisfait $x_i \geq 0$, $y_i \geq 0$, $z_i \geq 0$. Elle est donc admissible et fournit un tableau de départ pour l'algorithme du simplexe appliqué à (5.3).
- La fonction objective $-\sum_i z_i$ est majorée par zéro parce que $z_i \geq 0$. Le problème auxiliaire admet donc une solution finie.
- Pour la solution ainsi trouvée on a deux possibilités:
  (a) $\sum_i z_i = 0$, c.-à-d. $z_i = 0$ pout tout i. Dans cette situation, les valeurs x_i et y_i satisfont les contraintes du problème original (5.2). On a donc trouvé une solution de base admissible pour (5.2).
  (b) $\sum_i z_i < 0$. Si le problème (5.2) avait un point (x, y) satisfaisant les contraintes, la solution de (5.3) aurait $z = 0$ comme solution. L'ensemble des points admissibles pour (5.2) est donc vide dans cette situation.

Exemple 5.3 Considérons de nouveau le problème de l'exemple 4.1 pour lequel le vecteur b possède une composante négative. Le problème auxiliaire devient alors

$$\begin{array}{l} -z_1 \to \max \\ y_1 + x_1 + x_2 = 8 \\ y_2 - z_1 - x_1 - x_2 = -5 \\ y_3 + x_1 = 4 \\ y_4 + x_2 = 7 \\ x_i \geq 0, \;\; y_i \geq 0, \;\; z_i \geq 0. \end{array}$$



Comme proposé tout à l'heure, nous prenons la base $B = \{y_1, z_1, y_3, y_4\}$ et nous appliquons l'algorithme du simplexe au problème auxiliaire. Ceci donne

	x_1	x_2	y_2	
y_1	1	1	0	8
z_1	1	1	−1	5
y_3	$\boxed{1}$	0	0	4
y_4	0	1	0	7
	1	1	−1	5

	y_3	x_2	y_2	
y_1	−1	1	0	4
z_1	−1	$\boxed{1}$	−1	1
x_1	1	0	0	4
y_4	0	1	0	7
	−1	1	−1	1

	y_3	z_1	y_2	
y_1	0	−1	1	3
x_2	−1	1	−1	1
x_1	1	0	0	4
y_4	1	−1	1	6
	0	−1	0	0

(5.4)

et les itérations correspondent au chemin indiqué dans la figure. Si on avait choisi dans la première itération le pivot dans la colonne de x_2, on aurait tout de suite trouvé la solution de base admissible qui correspond à $x_1 = 0$ et $x_2 = 5$.

La valeur de la fonction objective est zéro (voir le dernier tableau de (5.4)). En supprimant la colonne correspondant à z_1, on obtient alors une solution de base admissible pour le problème original. Il faut encore adapter la fonction objective

$$-5x_1 - 4x_2 = -5(-y_3 + 4) - 4(y_3 + y_2 + 1) = y_3 - 4y_2 - 24$$

pour obtenir le tableau de départ cherché. L'algorithme du simplexe donne maintenant la solution en une itération:

	y_3	y_2	
y_1	0	1	3
x_2	−1	−1	1
x_1	$\boxed{1}$	0	4
y_4	1	1	6
	1	−4	24

	x_1	y_2	
y_1	0	1	3
x_2	1	−1	5
y_3	1	0	4
y_4	−1	1	2
	−1	−4	20

IV.6 Exercices

Chapter V

Calcul intégral

L'intégral d'une fonction a été introduit et discuté dans le cours "Analyse I" (semestre d'hiver). Le but de ce chapitre est d'exercer le calcul des intégrals et de donner des applications interessantes. Nous suivrons de près la présentation des paragraphes II.4 et II.5 du livre "L'analyse au fil de l'histoire" de Hairer & Wanner (Springer-Verlag 2000).

V.1 Primitives

Newton, Leibniz et Joh. Bernoulli découvrent indépendamment que l'intégration est une opération *inverse* de la différentiation.

Pour une fonction donnée $y = f(x)$, nous désignons par $z = F(x)$ l'aire sous $f(x)$ entre a et x (figure V.1, gauche). Le point crucial est que la fonction $f(x)$ est la dérivée de $F(x)$. Nous appelons alors $F(x)$ une *primitive* de $f(x)$ et nous la notons par $\int f(x)\,dx$. [1] On a alors

$$F(x) = \int f(x)\,dx + C \qquad \Longleftrightarrow \qquad F'(x) = f(x).$$

Pour Leibniz, l'aire en question est la *somme* (plus tard : "intégrale") des aires de rectangles infinitésimaux (figure V.1, droite).

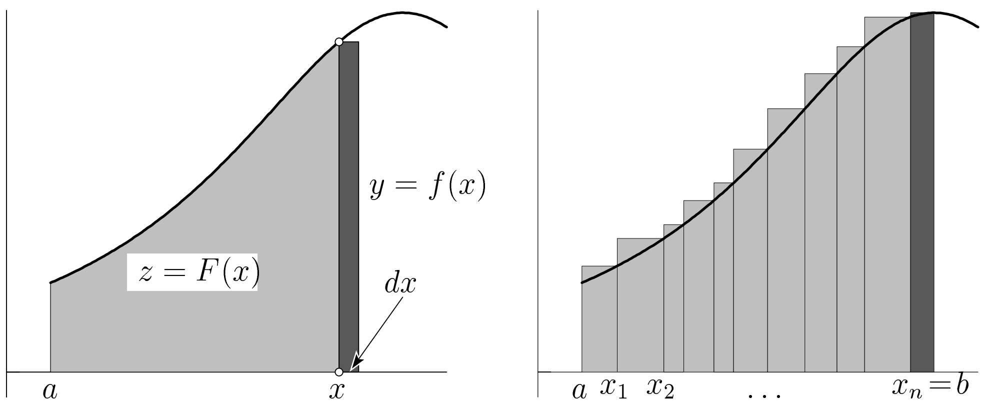


Figure V.1: Idée de Newton (gauche), idée de Leibniz (droite).

Les primitives ne sont pas uniques : à chaque primitive $F(x)$ peut être ajoutée une constante arbitraire C, car $F(x) + C$ est à nouveau une primitive de la même fonction. Pour $C = -F(a)$,

[1] Leibniz (1686) introduit le signe intégral, le mot "intégrale" est de Joh. Bernoulli, et fut publié par son frère Jac. Bernoulli (1690). La notation (1.1) pour l'aire entre les bornes a et b est due à Fourier (1822).

nous obtenons *la* primitive $F(x) - F(a)$ qui s'annule en $x = a$ (en même temps que l'aire z). Nous avons donc pour l'aire entre a et b

$$\int_a^b f(x)\,dx = F(b) - F(a) \tag{1.1}$$

(nous rappellons que cette identité est le "théorème fondamental du calcul différentiel", voir le cours "Analyse I").

Chaque formule de différentiation, inversée, livre une primitive. Par exemple, la dérivée de la fonction $f(x) = x^{n+1}$ est $f'(x) = (n+1)x^n$. Donc, $x^{n+1}/(n+1)$ est une primitive de x^n. La table suivante inverse de cette manière des formules bienconnues de la différentiation.

Petit tableau de primitives

$$\int x^n\,dx = \frac{x^{n+1}}{n+1} + C \quad (n \neq -1) \qquad \int \frac{1}{x}\,dx = \ln x + C$$

$$\int e^x\,dx = e^x + C$$

$$\int \sin x\,dx = -\cos x + C \qquad \int \cos x\,dx = \sin x + C$$

$$\int \frac{1}{1+x^2}\,dx = \arctan x + C \qquad \int \frac{1}{\sqrt{1-x^2}}\,dx = \arcsin x + C$$

Pour être utile, une table de primitives doit comporter plusieurs centaines de pages. Mentionnons ici celles de Gröbner & Hofreiter (1949) et de Gradshteyn & Ryzhik (1980). Aujourd'hui, de nombreux logiciels de calcul symbolique contiennent de telles tables.

V.2 Applications du calcul intégral

Aires de paraboles. L'aire sous la parabole $y = x^n$ de degré n entre a et b se calcule avec (1.1) et le tableau précédent :

$$\int_a^b x^n\,dx = \frac{x^{n+1}}{n+1}\Big|_a^b = \frac{b^{n+1} - a^{n+1}}{n+1}. \tag{2.1}$$

Nous avons utilisé la notation $F(x)|_a^b = F(b) - F(a)$.

Aire d'un disque. Pour calculer l'aire d'un quart de disque, prenons la fonction $f(x) = \sqrt{1-x^2}$ pour $0 \leq x \leq 1$. Une primitive de $f(x)$ est

$$F(x) = \frac{x}{2}\sqrt{1-x^2} + \frac{1}{2}\arcsin x, \tag{2.2}$$

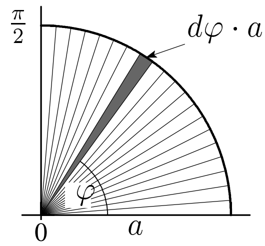
comme on peut le vérifier par différentiation. Nous verrons plus tard comment trouver de telles formules. Par (1.1), nous avons

$$\text{aire du disque unité } = 4\int_0^1 \sqrt{1-x^2}\,dx = 4\Big(F(1) - F(0)\Big) = \pi,$$

car $\sin(\pi/2) = 1$.

Voici un découpage plus élégant du disque. Rien ne nous oblige à supposer que $f(x)\,dx$ s'obtienne en découpant l'aire considérée en rectangles verticaux. Découpons le cercle (de rayon a) en triangles infiniment minces. L'aire d'un tel triangle est

$$dS = \frac{a^2 \cdot d\varphi}{2},$$



où $d\varphi$ est l'incrément infinitésimal de l'angle. L'aire totale (la somme des aires de tous ces triangles) est

$$S = \int_0^{2\pi} \frac{a^2\,d\varphi}{2} = \frac{a^2}{2}\int_0^{2\pi} d\varphi = \frac{a^2}{2}\,\varphi\Big|_0^{2\pi} = a^2\pi.$$

Volume de la boule. Considérons une boule de rayon a (voir la figure V.2) et découpons-la en tranches d'épaisseur dx et de rayon $r = \sqrt{a^2 - x^2}$. Le volume d'une telle tranche est $dV = r^2\pi\,dx = (a^2 - x^2)\pi\,dx$ et le volume total de la boule est

$$V = \int_{-a}^{+a} (a^2 - x^2)\pi\,dx = \pi\Big(xa^2 - \frac{x^3}{3}\Big)\Big|_{-a}^{+a} = \frac{4a^3\pi}{3}.$$

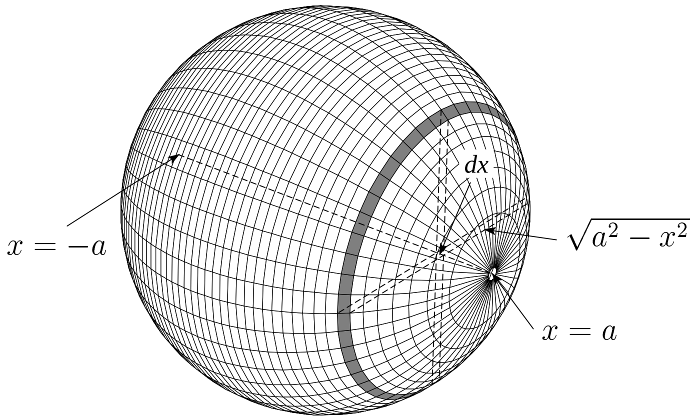
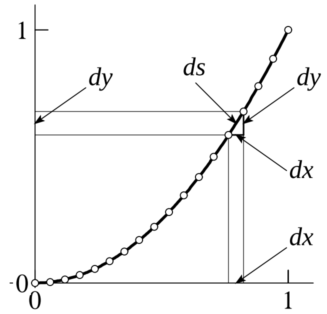


Figure V.2: Volume d'une boule

Longueur d'arc. On veut calculer la longueur d'arc L d'une courbe $y(x)$ donnée, entre $x = a$ et $x = b$. Si on augmente x de dx (voir la figure), l'ordonnée croît de $dy = y'(x)\,dx$ (en négligeant les termes d'ordre supérieur). La longueur d'une petite portion de la courbe est par conséquent donnée par ds, où



$$ds^2 = dx^2 + dy^2 = (1 + y'(x)^2)\,dx^2$$

(théorème de Pythagore). Si on fait tendre $dx \to 0$, on a

$$ds = \sqrt{1 + y'(x)^2} \cdot dx \qquad \text{et} \qquad L = \int_a^b \sqrt{1 + y'(x)^2}\;dx. \tag{2.3}$$

Exemple. Pour la parabole $y = x^2$, nous avons $y' = 2x$ et la longueur d'arc entre $x = 0$ et $x = 1$ est donnée par (voir la formule (3.15) plus loin)

$$L = \int_0^1 \sqrt{1 + 4x^2}\,dx = \frac{1}{2}\,x\sqrt{1 + 4x^2} + \frac{1}{4}\,\ln\Big(2x + \sqrt{4x^2 + 1}\Big)\Big|_0^1 = \frac{\sqrt{5}}{2} + \frac{1}{4}\,\ln(2 + \sqrt{5}\,).$$

V.3 Techniques d'intégration

Nous abordons ici des techniques générales permettant de trouver des primitives. Observons tout d'abord que l'intégration est une opération linéaire, c.-à-d.

$$\int \Big(c_1 f_1(x) + c_2 f_2(x) \Big)\, dx = c_1 \int f_1(x)\, dx + c_2 \int f_2(x)\, dx, \tag{3.1}$$

puisque la différentiation est une opération linéaire.

Substitution d'une nouvelle variable. Soit

$$F(z) \text{ une primitive de } f(z),$$

i.e. $F'(z) = f(z)$, et considérons la substitution $z = g(x)$ qui transforme la variable z en x. La formula pour la différentiation d'une fonction composée donne alors que

$$F(g(x)) \text{ est une primitive de } f(g(x))g'(x).$$

Par conséquent, nous avons

$$\int_a^b f(g(x))g'(x)\, dx = \int_{g(a)}^{g(b)} f(z)\, dz, \tag{3.2}$$

car, par (1.1) chaque membre est égal à $F(g(b)) - F(g(a))$. Le membre de gauche a été obtenu par la substitution $z = g(x)$ en $f(z)$ et avec $dz = g'(x)dx$.

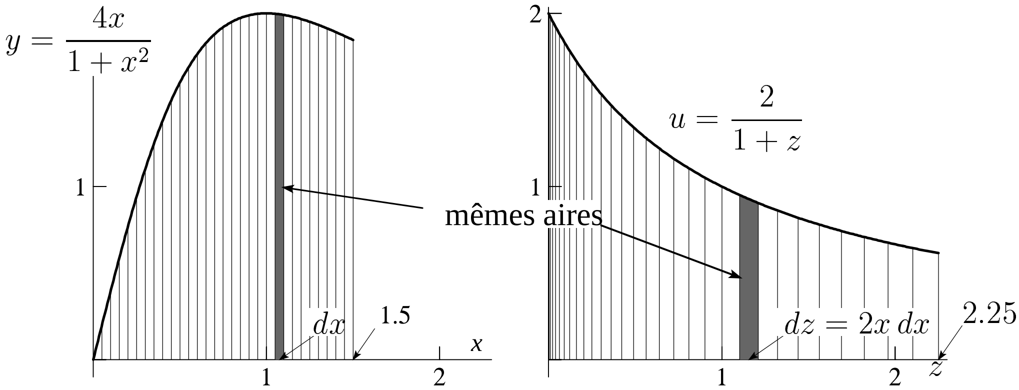


Figure V.3: Changement de variable dans une intégrale

Interprétation géométrique. Calculons

$$\int_0^{1.5} \frac{4x}{1+x^2}\, dx$$

avec la substitution $z = x^2$. Puisque $dz = 2x\,dx$, nous obtenons de (3.2)

$$\int_0^{1.5} \frac{2}{1+x^2} \cdot 2x\, dx = \int_0^{2.25} \frac{2}{1+z}\, dz = 2 \cdot \ln(1+z)\Big|_0^{2.25} = 2 \cdot \ln(1+x^2)\Big|_0^{1.5} = 2 \cdot \ln(3.25).$$

La figure V.3 illustre le changement de variable $z = x^2$ transformant la fonction $4x/(1+x^2)$ en $2/(1+z)$. Les points x et $x + dx$ sont tranformés en $z = x^2$ et $z + dz = x^2 + 2x\,dx + dx^2$. Pour

$dx \to 0$, les rectangles en noir ont les mêmes aires et par conséquent, les deux intégrales de (3.2) ont la même valeur.

Exemples. L'art consiste à trouver de “bonnes” substitutions. Démontrons-le dans une suite d'exemples.

Pour des fonctions de la forme $f(ax+b)$, on pose $z = ax+b$. Par exemple, avec $z = 5x+2$, $dz = 5\,dx$, nous avons (en omettant la constante d'intégration)

$$\int e^{5x+2}\,dx = \int e^z\,\frac{dz}{5} = \frac{1}{5}\,e^z = \frac{1}{5}\,e^{5x+2}. \tag{3.3}$$

Parfois, le facteur $g'(x)$ est facilement reconnaissable pour la substitution $z = g(x)$. Dans l'intégrale suivante, par exemple, le facteur x incite à poser $z = -x^2$, $dz = -2x\,dx$. On obtient alors

$$\int xe^{-x^2}\,dx = -\frac{1}{2}\int e^z\,dz = -\frac{1}{2}\,e^z = -\frac{1}{2}\,e^{-x^2}. \tag{3.4}$$

Le tableau du paragraphe V.1 donne les intégrales de $1/(1+x^2)$ ou de $1/\sqrt{1-x^2}$. Pour trouver la primitive de $1/(7+x^2)$ ou de $1/\sqrt{7-x^2}$, on utilise la substitution $x^2 = 7z^2$ ou $x = \sqrt{7}\,z$, $dx = \sqrt{7}\,dz$. Cela donne par exemple,

$$\int \frac{dx}{7+x^2} = \int \frac{\sqrt{7}\,dz}{7(1+z^2)} = \frac{1}{\sqrt{7}}\arctan z = \frac{1}{\sqrt{7}}\arctan\frac{x}{\sqrt{7}}. \tag{3.5}$$

Les expressions quadratiques $x^2+2bx+c$ sont souvent simplifiées en “complétant le carré” $x^2+2bx+c = (x+b)^2+(c-b^2)$, puis en faisant la substitution $z = x+b$. Ainsi, l'intégrale suivante est réduite, par la substitution $z = x+1/2$, à l'intégrale (3.5) :

$$\int \frac{dx}{x^2+x+1} = \int \frac{dz}{z^2+3/4} = \frac{2}{\sqrt{3}}\arctan\frac{2z}{\sqrt{3}} = \frac{2}{\sqrt{3}}\arctan\Big(\frac{2x+1}{\sqrt{3}}\Big). \tag{3.6}$$

Comme dernier exemple, prenons la fonction $(x+2)/(x^2+x+1)$. Nous écrivons le numérateur sous la forme $x+2 = (x+1/2)+3/2$, pour que la première partie $x+1/2$ soit un multiple de la dérivée du dénominateur. Cette partie de l'intégrale est ensuite calculée par la substitution $z = x^2+x+1$. La deuxième partie est un multiple de (3.6), et nous avons

$$\int \frac{x+2}{x^2+x+1}\,dx = \frac{1}{2}\ln(x^2+x+1) + \sqrt{3}\arctan\Big(\frac{2x+1}{\sqrt{3}}\Big). \tag{3.7}$$

Intégration par parties. Une deuxième technique d'intégration est basée sur la règle de différentiation d'un produit de deux fonctions. La formule $(uv)' = u'v+uv'$ donne par intégration $u(x)v(x) = \int (u'(x)v(x)+u(x)v'(x))\,dx$, d'où

$$\int u'(x)v(x)\,dx = u(x)v(x) - \int u(x)v'(x)\,dx. \tag{3.8}$$

Cette formule ramène le calcul d'une intégrale à celui d'une autre. Si les facteurs u' et v sont bien choisis, cette deuxième intégrale sera plus facile à calculer que la première.

Exemples. Calculons $\int x\sin x\,dx$. Il ne serait pas judicieux de choisir $u'(x) = x$ (c.-à-d. $u(x) = x^2/2$) et $v(x) = \sin x$, car la deuxième intégrale serait encore plus difficile. Choisissons donc $u'(x) = \sin x$ (donc $u(x) = -\cos x$) et $v(x) = x$. La formule (3.8) donne alors

$$\int x\sin x\,dx = -x\cos x + \int 1\cdot\cos x\,dx = -x\cos x + \sin x. \tag{3.9}$$

Il arrive que l'intégration par parties doive être répétée. Dans l'exemple suivant, nous posons d'abord $v(x) = x^2$, $u'(x) = e^x$, puis faisons une deuxième intégration par parties avec $v(x) = x$, $u'(x) = e^x$:

$$\int x^2\, e^x\, dx = x^2 e^x - 2\int x\, e^x\, dx = e^x(x^2 - 2x + 2). \tag{3.10}$$

Des fonctions telles que $\ln(x)$, $\arctan(x)$ ont des dérivées simples ; on les utilisera souvent pour $v(x)$:

$$\int \ln x\, dx = \int 1 \cdot \ln x\, dx = x \ln x - \int \frac{x}{x}\, dx = x(\ln x - 1), \tag{3.11}$$

$$\int \arctan x\, dx = x \arctan x - \int \frac{x}{1+x^2}\, dx = x \arctan x - \frac{1}{2}\ln(1+x^2)\ ; \tag{3.12}$$

la dernière intégrale a été calculée avec la substitution $z = 1 + x^2$, $dz = 2x\, dx$.

Pour calculer l'intégrale $\int \sqrt{1+4x^2}\, dx$ (rencontrée dans le calcul de la longueur d'arc d'une parabole), nous faisons une intégration par parties avec $u'(x) = 1$, $v(x) = \sqrt{1+4x^2}$:

$$\int \sqrt{1+4x^2}\, dx = x\sqrt{1+4x^2} - \int \frac{4x^2}{\sqrt{1+4x^2}}\, dx. \tag{3.13}$$

Si nous écrivons le numérateur de la dernière intégrale sous la forme $4x^2 = (1+4x^2) - 1$, l'intégrale se décompose en deux intégrales : l'intégrale cherchée $-\int \sqrt{1+4x^2}\, dx$ et la deuxième intégrale semblable à la dernière du tableau du paragraphe V.1 : la dérivée de $\operatorname{arsinh} z$ est $1/\sqrt{1+z^2}$ et la substitution $z = 2x$ donne

$$\int \frac{dx}{\sqrt{1+4x^2}} = \frac{1}{2}\operatorname{arsinh}(2x) = \frac{1}{2}\ln\Big(2x + \sqrt{4x^2+1}\Big). \tag{3.14}$$

Ainsi (3.13) livre

$$\int \sqrt{1+4x^2}\, dx = \frac{1}{2}x\sqrt{1+4x^2} + \frac{1}{4}\ln\Big(2x + \sqrt{4x^2+1}\Big). \tag{3.15}$$

Relations de récurrence. Pour calculer l'intégrale

$$J_n = \int \frac{dx}{(1+x^2)^n}, \tag{3.16}$$

intégrons par parties avec $u'(x) = 1$ et $v(x) = 1/(1+x^2)^n$:

$$J_n = \int 1 \cdot \frac{1}{(1+x^2)^n}\, dx = \frac{x}{(1+x^2)^n} + n\int \frac{2x^2}{(1+x^2)^{n+1}}\, dx$$

et, comme en (3.13), écrivons dans la dernière intégrale $2x^2 = 2(1+x^2) - 2$ pour obtenir

$$J_n = \frac{x}{(1+x^2)^n} + 2nJ_n - 2nJ_{n+1}.$$

Il semble que notre choix ait été maladroit : au lieu de décroître, l'indice n est devenu *plus grand*. Renversons simplement la formule :

$$J_{n+1} = \frac{1}{2n}\,\frac{x}{(1+x^2)^n} + \frac{2n-1}{2n}\,J_n. \tag{3.17}$$

Cette relation ramène le calcul de (3.16) à celui de l'intégrale $J_1 = \arctan x$.

V.4 Intégration de fonctions rationnelles

Soit $R(x) = P(x)/Q(x)$ une fonction rationnelle ($P(x)$ et $Q(x)$ des polynômes). Si on sait calculer les zéros de $Q(x)$, nous allons voir qu'on peut toujours trouver une primitive $\int R(x)\,dx$. Elle sera trouvée en trois étapes :

– réduction au cas $\deg P < \deg Q$ (où $\deg P$ est le degré de $P(x)$) ;
– factorisation de $Q(x)$ et décomposition de $R(x)$ en fractions simples ;
– intégration des fractions simples.

Réduction au cas $\deg P < \deg Q$**.** On effectue une première simplification de la fonction $R(x)$, si $\deg P \geq \deg Q$. Dans cette situation, on divise P par Q pour trouver

$$\frac{P(x)}{Q(x)} = S(x) + \frac{\widehat{P}(x)}{Q(x)}, \tag{4.1}$$

où $S(x)$ et $\widehat{P}(x)$ sont des polynômes (quotient et reste) avec $\deg \widehat{P} < \deg Q$. Prenons pour exemple

$$\frac{P(x)}{Q(x)} = \frac{2x^6 - 3x^5 - 9x^4 + 23x^3 + x^2 - 44x + 39}{x^5 + x^4 - 5x^3 - x^2 + 8x - 4}. \tag{4.2}$$

Tout d'abord, on fait disparaître le terme $2x^6$ en retranchant $2xQ(x)$ de $P(x)$, puis on additionne $5Q(x)$ à $P(x)$ pour obtenir

$$\frac{P(x)}{Q(x)} = 2x - 5 + \frac{6x^4 - 20x^2 + 4x + 19}{x^5 + x^4 - 5x^3 - x^2 + 8x - 4}. \tag{4.3}$$

Le polynôme $S(x)$ est facilement intégré ; seul le deuxième terme dans (4.1) nécessite des efforts supplémentaires.

Décomposition en fractions simples. Supposons qu'on connaisse les racines de $Q(x)$. Notons les racines réelles distinctes par $\gamma_1, \ldots, \gamma_k$ et les racines complexes conjuguées distinctes par $\alpha_1 \pm i\beta_1, \ldots, \alpha_l \pm i\beta_l$ (qui apparaissent en paires, si le polynôme $Q(x)$ possède des coefficients réelles). Dans ce cas, nous obtenons la factorisation en polynômes réels

$$Q(x) = \prod_{i=1}^{k} (x - \gamma_i)^{n_i} \prod_{i=1}^{l} ((x - \alpha_i)^2 + \beta_i^2)^{m_i}, \tag{4.4}$$

où les m_i et n_i sont les multiplicités des racines. Le lemme suivant montre comment une fonction rationnelle peut être décomposée en somme de *fractions simples*.

Lemme 4.1 *Soit $Q(x)$ donné par (4.4) et $P(x)$ un polynôme à coefficients réels avec $\deg P < \deg Q$. Alors il existe des constantes réelles A_{ij}, B_{ij} et C_{ij} telles que*

$$\frac{P(x)}{Q(x)} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{C_{ij}}{(x - \gamma_i)^j} + \sum_{i=1}^{l} \sum_{j=1}^{m_i} \frac{A_{ij} + B_{ij}x}{((x - \alpha_i)^2 + \beta_i^2)^j}. \tag{4.5}$$

Démonstration. Nous traitons les facteurs de $Q(x)$ l'un après l'autre. Commenons par les racines réelles et, pour cela, nous posons $Q(x) = (x - \gamma)^n q(x)$, où γ est une racine de $Q(x)$ et $q(\gamma) \neq 0$. Démontrons maintenant l'existence d'une constante C et d'un polynôme $p(x)$ de degré $< \deg Q - 1$ tels que

$$\frac{P(x)}{(x - \gamma)^n q(x)} = \frac{C}{(x - \gamma)^n} + \frac{p(x)}{(x - \gamma)^{n-1} q(x)}, \tag{4.6}$$

c'est-à-dire, après multiplication par le dénominateur commun,

$$P(x) = C \cdot q(x) + p(x) \cdot (x - \gamma). \tag{4.7}$$

En posant $x = \gamma$, on voit que $C = P(\gamma)/q(\gamma)$. Pour obtenir le polynôme $p(x)$, nous divisons $P(x) - C \cdot q(x)$ par $(x - \gamma)$. Appliquons ensuite récursivement le même procédé au second terme du membre de droite de (4.6). Nous parvenons alors à la première partie de la décomposition (4.5) annoncée.

Quant aux racines complexes, nous écrivons $Q(x) = ((x-\alpha)^2 + \beta^2)^m q(x)$, où $\alpha + i\beta$ est une racine de $Q(x)$ et $q(\alpha \pm i\beta) \neq 0$. Alors il existe des constantes réelles A, B et un polynôme $p(x)$ de degré $< \deg Q - 2$ tels que

$$\frac{P(x)}{((x-\alpha)^2 + \beta^2)^m q(x)} = \frac{A + Bx}{((x-\alpha)^2 + \beta^2)^m} + \frac{p(x)}{((x-\alpha)^2 + \beta^2)^{m-1} q(x)}.$$

Pour le voir, nous prouvons l'équation équivalente

$$P(x) = (A + Bx) \cdot q(x) + p(x) \cdot ((x-\alpha)^2 + \beta^2).$$

En posant $x = \alpha + i\beta$, nous déterminons A et B en comparant les parties réelles et imaginaires. Le polynôme $p(x)$ est alors obtenu en divisant $P(x) - (A + Bx) \cdot q(x)$ par $((x-\alpha)^2 + \beta^2)$. Nous parvenons donc à la formule (4.5) par récurrence sur le degré de $Q(x)$. □

Exemple. Le polynôme $Q(x)$ de (4.2) admet la factorisation

$$Q(x) = x^5 + x^4 - 5x^3 - x^2 + 8x - 4 = (x-1)^3(x+2)^2. \tag{4.8}$$

On applique (4.7) en posant $\gamma = -2$ et $n = 2$, et (4.6) livre

$$\frac{6x^4 - 20x^2 + 4x + 19}{(x-1)^3(x+2)^2} = \frac{-1}{(x+2)^2} + \frac{6x^3 - 11x^2 - x + 9}{(x-1)^3(x+2)}.$$

Deuxième étape : posons $\gamma = -2$ et $n = 1$ et nous avons

$$\frac{6x^4 - 20x^2 + 4x + 19}{(x-1)^3(x+2)^2} = \frac{-1}{(x+2)^2} + \frac{3}{x+2} + \frac{3x^2 - 8x + 6}{(x-1)^3}. \tag{4.9}$$

Dans la dernière expression, nous écrivons $x = (x-1) + 1$ pour que $3x^2 - 8x + 6 = 3(x-1)^2 - 2(x-1) + 1$, et (4.9) devient enfin

$$\frac{6x^4 - 20x^2 + 4x + 19}{(x-1)^3(x+2)^2} = \frac{1}{(x-1)^3} + \frac{-2}{(x-1)^2} + \frac{3}{x-1} + \frac{-1}{(x+2)^2} + \frac{3}{x+2}. \tag{4.10}$$

Deuxième possibilité. Par le lemme 4.1, nous pouvons écrire

$$\frac{6x^4 - 20x^2 + 4x + 19}{(x-1)^3(x+2)^2} = \frac{A_0}{(x-1)^3} + \frac{A_1}{(x-1)^2} + \frac{A_2}{x-1} + \frac{B_0}{(x+2)^2} + \frac{B_1}{x+2}. \tag{4.11}$$

On calcule les coefficients A_i en multipliant (4.11) par $(x-1)^3$:

$$\frac{6x^4 - 20x^2 + 4x + 19}{(x+2)^2} = A_0 + A_1(x-1) + A_2(x-1)^2 + (x-1)^3 g(x),$$

où $g(x)$ est une fonction bien définie au voisinage de $x = 1$. Les A_i sont donc les premiers coefficients de la série de Taylor de $P(x)/(x+2)^2$, donnés par

$$A_i = \frac{1}{i!}\frac{d^i}{dx^i}\Big(\frac{6x^4-20x^2+4x+19}{(x+2)^2}\Big)\Big|_{x=1},$$

i.e. $A_0 = 1$, $A_1 = -2$, $A_2 = 3$. De manière analogue, nous obtenons

$$B_i = \frac{1}{i!}\frac{d^i}{dx^i}\Big(\frac{6x^4-20x^2+4x+19}{(x-1)^3}\Big)\Big|_{x=-2},$$

i.e. $B_0 = -1$, $B_1 = 3$.

Intégration de fractions simples. Dans la troisième étape, les termes de la première somme dans (4.5) sont intégrés en utilisant les formules du paragraphe V.1 (cf. le petit tableau) :

$$\int \frac{dx}{(x-\gamma)^j} = \begin{cases} \dfrac{-1}{(j-1)(x-\gamma)^{j-1}} & \text{si } j>1 \\ \ln(x-\gamma) & \text{si } j=1. \end{cases} \tag{4.12}$$

Pour l'exemple (4.2), nous avons avec (4.3), (4.8) et (4.10),

$$\int \frac{P(x)}{Q(x)}\,dx = x^2-5x-\frac{1}{2(x-1)^2}+\frac{2}{x-1}+3\ln(x-1)+\frac{1}{x+2}+3\ln(x+2)+C.$$

Si toutes les racines de $Q(x)$ sont réelles, nous avons alors trouvé une primitive de $P(x)/Q(x)$.

Pour intégrer un terme général de la deuxième expression du membre de droite dans l'équation (4.5), nous l'écrivons sous la forme

$$\frac{A+Bx}{((x-\alpha)^2+\beta^2)^j} = \frac{B(x-\alpha)}{((x-\alpha)^2+\beta^2)^j} + \frac{A+B\alpha}{((x-\alpha)^2+\beta^2)^j}.$$

Le premier terme de cette expression est intégré en faisant la substitution $z = (x-\alpha)^2+\beta^2$, $dz = 2(x-\alpha)dx$. Pour le deuxième terme, nous utilisons la substitution $z = (x-\alpha)/\beta$ et obtenons l'intégrale (3.16) du paragraphe V.3. Ainsi, nous avons pour $j = 1$

$$\int \frac{A+Bx}{(x-\alpha)^2+\beta^2}\,dx = \frac{B}{2}\,\ln((x-\alpha)^2+\beta^2) + \frac{A+B\alpha}{\beta}\arctan\Big(\frac{x-\alpha}{\beta}\Big),$$

et pour $j > 1$

$$\int \frac{A+Bx}{((x-\alpha)^2+\beta^2)^j}\,dx = \frac{-B}{2(j-1)((x-\alpha)^2+\beta^2)^{j-1}} + \frac{A+B\alpha}{\beta^{2j-1}}J_j\Big(\frac{x-\alpha}{\beta}\Big),$$

où $J_1(z) = \arctan z$ et par (3.17)

$$J_{j+1}(z) = \frac{z}{2j(z^2+1)^j} + \frac{2j-1}{2j}J_j(z). \tag{4.13}$$

Exemple. Le lemme 4.1 donne la décomposition suivante pour la fonction $(1+x^4)^{-1}$:

$$\frac{1}{1+x^4} = \frac{1}{(x^2+\sqrt{2}x+1)(x^2-\sqrt{2}x+1)} = \frac{A+Bx}{x^2+\sqrt{2}x+1} + \frac{C+Dx}{x^2-\sqrt{2}x+1}.$$

Multiplions cette relation par $(x^2 + \sqrt{2}x + 1)$ et posons $x = (-1 \pm i)/\sqrt{2}$. Cela donne

$$\frac{1}{2 \mp 2i} = \frac{1 \pm i}{4} = A + B\frac{(-1 \pm i)}{\sqrt{2}},$$

et $A = 1/2$, $B = \sqrt{2}/4$ apparaissent en comparant les parties réelles et imaginaires des deux membres. Les constantes $C = 1/2$ et $D = -\sqrt{2}/4$ s'obtiennent de manière analogue. Ces calculs livrent finalement le résultat

$$\int \frac{dx}{1+x^4} = \frac{\sqrt{2}}{8} \ln \frac{x^2 + \sqrt{2}x + 1}{x^2 - \sqrt{2}x + 1} + \frac{\sqrt{2}}{4}\Big(\arctan(x\sqrt{2}+1) + \arctan(x\sqrt{2}-1)\Big). \tag{4.14}$$

Remarque. La décomposition en fractions simples fut un des stimulants au réveil de l'intérêt des mathématiciens du XVIIIe siècle pour les racines des polynômes et pour l'algèbre.

V.5 Substitutions importantes

Le résultat précédent va nous permettre, grâce à diverses substitutions, de trouver la primitive pour d'autres classes de fonctions. Dans tout ce qui suit, R représente une fonction rationnelle à une, deux ou trois arguments.

Intégrales de la forme $\int R(\sqrt[n]{ax+b}, x)dx$. Voici une substitution évidente :

$$\sqrt[n]{ax+b} = u, \qquad x = \frac{u^n - b}{a}, \qquad dx = \frac{n}{a} \cdot u^{n-1} \cdot du, \tag{5.1}$$

elle donne

$$\int R\Big(\sqrt[n]{ax+b}, x\Big)\, dx = \frac{n}{a} \int R\Big(u, \frac{u^n - b}{a}\Big) u^{n-1}\, du = \int \widetilde{R}(u)\, du,$$

où $\widetilde{R}(u)$ est une fonction rationnelle. On peut donc calculer cette dernière intégrale par les techniques développées ci-dessus.

Intégrales de la forme $\int R(e^{\lambda x})dx$. Il est évident de poser ici $u = e^{\lambda x}$ avec $du = \lambda e^{\lambda x} dx$ et $dx = du/(\lambda u)$. La fonction à intégrer qui en résulte est une fonction rationnelle.

Exemple,

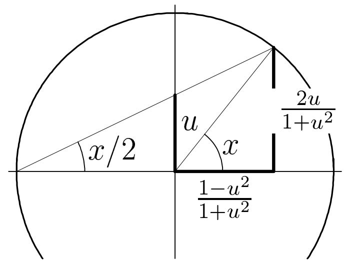
$$\begin{aligned}\int \frac{dx}{2 + \sinh x} &= \int \frac{dx}{2 + (e^x - e^{-x})/2} = 2\int \frac{du}{u^2 + 4u - 1} = \\ &= 2\int \frac{du}{(u+2)^2 - 5} = \frac{1}{\sqrt{5}} \ln \frac{u + 2 - \sqrt{5}}{u + 2 + \sqrt{5}} = \frac{1}{\sqrt{5}} \ln \frac{e^x + 2 - \sqrt{5}}{e^x + 2 + \sqrt{5}}.\end{aligned}$$

Intégrales de la forme $\int R(\sin x, \cos x, \tan x)dx$. Une des plus célèbres découvertes de l'Antiquité grecque (Pythagore 570–501 av. J.-C., voir aussi R.C. Buck 1980, *Sherlock Holmes in Babylon*, Am. Math. Monthly vol. 87, Nr. 5, p. 335-345) sont les triplets $(3, 4, 5)$, $(5, 12, 13)$, $(7, 24, 25)$, $\ldots$, qui satisfont $a^2 + b^2 = c^2$ et sont de la forme $(u, (u^2-1)/2, (u^2+1)/2)$; cela suggère la substitution

$$\sin x = \frac{2u}{1+u^2}, \qquad \cos x = \frac{1-u^2}{1+u^2}, \qquad \tan x = \frac{2u}{1-u^2}. \tag{5.2}$$

On vérifie que $\sin x = u(1+\cos x)$; par conséquent le point $(\cos x, \sin x)$ est l'intersection de la droite $\eta = u(1+\xi)$ et du cercle unité (voir la figure). Donc $u = \tan(x/2)$, $x = 2\arctan u$, et

$$dx = \frac{2}{1+u^2}\,du.$$



Cette substitution transforme $\int R(\sin x, \cos x, \tan x)dx$ en intégrale d'une fonction rationnelle.

Exemple.

$$\int \frac{dx}{2+\sin x} = \int \frac{2du}{(1+u^2)(2+\frac{2u}{1+u^2})} = \int \frac{du}{u^2+u+1}.$$

La dernière intégrale nous est déjà connue (4.18) ; par conséquent,

$$\int \frac{dx}{2+\sin x} = \frac{2}{\sqrt{3}}\,\arctan\Big(\frac{2}{\sqrt{3}}\Big(u+\frac{1}{2}\Big)\Big) = \frac{2}{\sqrt{3}}\,\arctan\Big(\frac{2}{\sqrt{3}}\Big(\tan\frac{x}{2}+\frac{1}{2}\Big)\Big).$$

Intégrales de la forme $\int R(\sqrt{ax^2+2bx+c}, x)dx$. L'idée (Euler 1768, § 88) est de définir une nouvelle variable z par la relation $ax^2+2bx+c = a(x-z)^2$. Cela donne la substitution

$$\begin{aligned} &x = \frac{az^2-c}{2(b+az)}, \qquad dx = \frac{a(az^2+2bz+c)}{2(b+az)^2}\,dz, \\ &\sqrt{ax^2+2bx+c} = \pm\sqrt{a}\,(z-x) = \pm\sqrt{a}\cdot\frac{az^2+2bz+c}{2(b+az)}, \\ &z = x \pm \frac{\sqrt{ax^2+2bx+c}}{\sqrt{a}}, \end{aligned} \tag{5.3}$$

et il s'agit à nouveau de calculer l'intégrale d'une fonction rationnelle.

Parfois, il est plus simple de transformer l'expression $\sqrt{ax^2+2bx+c}$ par une substitution linéaire $z = \alpha x + \beta$, en une des formes

$$\sqrt{z^2+1}, \qquad \sqrt{z^2-1}, \qquad \sqrt{1-z^2}.$$

Puis, les substitutions

$$z = \sinh u, \qquad z = \cosh u, \qquad z = \sin u \tag{5.4}$$

permettent d'éliminer les racines carrées de l'intégrale.

Exemple. Considérons, une fois de plus, l'intégrale (3.15). En posant $x = \sinh u$, nous obtenons

$$\begin{aligned} \int \sqrt{x^2+1}\,dx &= \int \cosh^2 u\,du = \int \Big(\frac{1}{2}+\frac{\cosh 2u}{2}\Big)\,du = \frac{u}{2}+\frac{\sinh 2u}{4} \\ &= \frac{u}{2}+\frac{\sinh u\cosh u}{2} = \frac{1}{2}\,\ln(x+\sqrt{x^2+1})+\frac{x\sqrt{x^2+1}}{2}. \end{aligned}$$

Remarquons, que la fonction inverse de $x = \sinh u$ est obtenue en posant $v = e^u$ dans la définition de $\sinh u$ et en calculant les racines du polynôme $v^2 - 2xv - 1 = 0$ ce qui est équivant à $2x = v - v^{-1}$.

V.6 Exercices

Chapter VI

Equations différentielles ordinaires

Une *équation différentielle ordinaire* est une équation de la forme

$$y' = f(x,y) \qquad \text{ou} \qquad y'' = f(x,y,y') \qquad \text{ou} \qquad \ldots$$

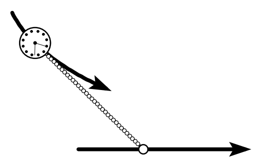
Il s'agit donc de trouver une fonction $y(x)$ telle que $y'(x) = f(x, y(x))$ ou $y''(x) = f(x, x(x), y'(x))$ pour tout x dans un certain intervalle.

Dans ce chapitre nous commençons par le cas où $y(x)$ est une fonction scalaire. Nous traitons quelques exemples historiques, et des problèmes qu'on peut résoudre analytiquement. (Cette partie suit de près la présentation des paragraphes II.7 et II.8 du livre "L'analyse au fil de l'histoire" de Hairer & Wanner.) Ensuite nous considérons la situation où $y(x)$ est une fonction vectorielle ($y : \mathbb{R} \to \mathbb{R}^n$) et nous étudions l'existence et l'unicité des solutions.

VI.1 Exemples historiques

Un excellent article sur l'origine des équations différentielles est écrit par G. Wanner[1]. Nous en prenons deux exemples.

VI.1.1 La tractrice



Lors du séjour de Leibniz à Paris (1672–1676) durant lequel il suit des cours d'Huygens, Claude Perrault, célèbre anatomiste et architecte, lui pose le problème suivant : *quelle est la courbe qui a la propriété qu'en chacun de ses points P le segment de la tangente entre P et l'axe x est de longueur constante a ?* Pour concrétiser cette question, Perrault tire de son gousset une "horologio portabili suae thecae argenteae" et la fait glisser sur la table. Il précise qu'aucun mathématicien parisien ni toulousain (Fermat) n'a été capable de trouver l'équation de la courbe.

Leibniz publie sa solution en 1693 en affirmant la connaître depuis longtemps : puisque

$$\frac{dy}{dx} = -\frac{y}{\sqrt{a^2 - y^2}}, \qquad \text{i.e.} \qquad -\frac{\sqrt{a^2-y^2}}{y}\, dy = dx, \tag{1.1}$$

on trouve la solution par quadratures. Leibniz affirme que c'est un "fait bien connu" que cette aire peut être exprimée à l'aide d'un logarithme, ce qui se vérifie avec la substitution $\sqrt{a^2 - y^2} = v$,

[1] G. Wanner, *Les équations différentielles ont 350 ans*, L'Enseignement Mathématique 34 (1988), 365–385.

$a^2 - y^2 = v^2$, $-y\,dy = v\,dv$, qui mène à

$$\begin{aligned} x + C &= -\int \frac{\sqrt{a^2-y^2}}{y}\,dy = \int \frac{v^2}{a^2-v^2}\,dv = \int\Big(-1 + \frac{a}{2}\Big(\frac{1}{a-v} + \frac{1}{a+v}\Big)\Big)dv \\ &= -v - \frac{a}{2}\log\frac{a-v}{a+v} = -\sqrt{a^2-y^2} - a\,\log\frac{a-\sqrt{a^2-y^2}}{y}. \end{aligned}$$

Si l'on veut que $y = a$ pour $x = 0$, la constante d'intégartion s'annule.

VI.1.2 La caténaire

Galilée (1638) affirme que la forme d'une chaîne suspendue entre deux clous est presque une parabole. Environ 20 ans plus tard, un jeune Hollandais âgé de 16 ans (Christiaan Huygens) démontre l'impossibilité de ce résultat. Enfin, la solution du problème de la forme d'une corde flexible suspendue ("Linea Catenaria vel Funicularis") par Leibniz (1691) et Joh. Bernoulli (1691) fut un succès considérable pour le calcul différentiel.

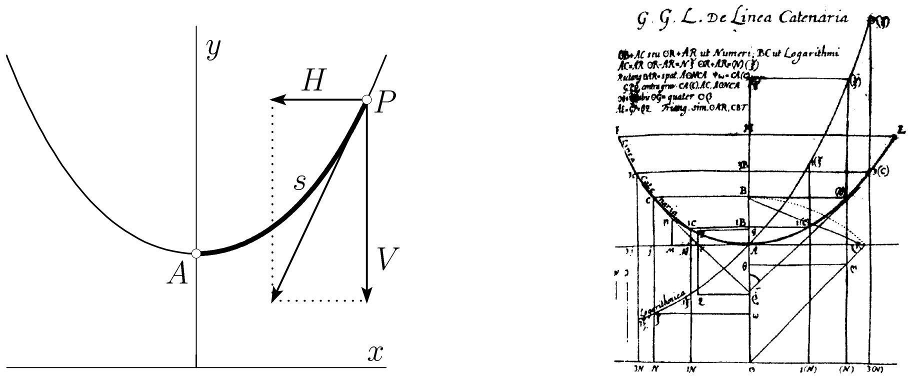


Figure VI.1: La caténaire (gauche) et un dessin de Leibniz de 1691 (droite)

Pour un point P de la courbe, considérons la force horizontale H et la force verticale V (figure VI.1). Nous avons alors que

$$H = a, \qquad V = q\,s$$

(où a et q sont des constantes et s est la longeur d'arc AP) parce qu'il n'y a pas de force horizontale extérieure et V représente la pesanteur de la partie AP de la chaîne. En notant la pente en P par $p = y'$, nous obtenons

$$p = \frac{dy}{dx} = \frac{V}{H} = \frac{qs}{a}\,.$$

Après différentiation, nous avons avec $c = a/q$

$$c \cdot dp = ds = \sqrt{1+p^2}\,dx, \tag{1.2}$$

une équation différentielle pour les variables p et x. En utilisant la substitution $p = \sinh u$, une intégration donne[2]

$$x - x_0 = c\int \frac{dp}{\sqrt{1+p^2}} = c\int du = c \cdot u.$$

Par conséquent,

$$p = \sinh\Big(\frac{x-x_0}{c}\Big) \qquad \text{et} \qquad y = K + c\cdot\cosh\Big(\frac{x-x_0}{c}\Big). \tag{1.3}$$

[2]Rappelons que $\sinh u = (e^u - e^{-u})/2$, $\cosh u = (e^u + e^{-u})/2$, $\cosh^2 u - \sinh^2 u = 1$, la dérivée de $\sinh u$ est $\cosh u$ et la dérivée de $\cosh u$ est $\sinh u$.

VI.2 Quelques types d'équations intégrables

Discutons maintenant de quelques types simples d'équations différentielles qui peuvent être résolues par des calculs d'intégrales ("par quadrature").

VI.2.1 Équations à variables séparées

Tous les exemples précédents, à savoir (1.1) et (1.2), sont du type

$$y' = f(x)g(y). \tag{2.1}$$

On les résout en écrivant $y' = dy/dx$, en "séparant les variables" et en intégrant, i.e.

$$\frac{dy}{g(y)} = f(x)\,dx \qquad \text{et} \qquad \int \frac{dy}{g(y)} = \int f(x)\,dx + C. \tag{2.2}$$

Si $G(y)$ et $F(x)$ sont des primitives de $1/g(y)$ et de $f(x)$ respectivement, la solution est exprimée par $G(y) = F(x) + C$.

Exemple 2.1 La loi logistique (Verhulst 1837), décrivant la croissance d'une population, est donné par l'équation différentielle

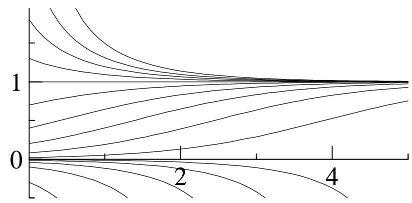
$$\dot{y} = ay - by^2$$

où a et b sont des constantes positives ($\dot{y}$ représente la dérivée par rapport au temps; nous écrivons t au lieu de x). Supposons que $a = b = 1$. En séparant les variables et en utilisant

$$\frac{dy}{y(1-y)} = \frac{dy}{y} + \frac{dy}{1-y},$$

on trouve la solution sous la forme (avec la constante déterminé par $y(0) = y_0$)

$$\ln\Big(\frac{y}{1-y}\Big) = t + \overline{C} \qquad \text{resp.} \qquad y(t) = \frac{Ce^t}{1+Ce^t}.$$



VI.2.2 Équations linéaires homogènes

Un cas particulier important de (2.1) est

$$y' = f(x)y. \tag{2.3}$$

Sa solution est donnée par

$$\ln y = \int f(x)\,dx + \overline{C} \qquad \text{ou} \qquad y = C \cdot \exp\Big(\int f(x)\,dx\Big). \tag{2.4}$$

VI.2.3 Équations linéaires inhomogènes

Elle est de la forme

$$y' = f(x)y + g(x) \tag{2.5}$$

et le terme $g(x)$ est l'inhomogénéité de l'équation différentielle. Le procédé de résolution s'appelle *variation des constantes*. L'idée est de prendre la solution (2.4) de l'équation homogène avec C remplaçée par une fonction $c(x)$ et de déterminer $c(x)$ afin de satisfaire (2.5). On pose

$$y(x) = c(x) \cdot u(x) \qquad \text{où} \qquad u(x) = \exp\Big(\int_0^x f(t)\,dt\Big), \tag{2.6}$$

on calcule sa dérivée $y'(x) = c'(x)u(x) + c(x)u'(x) = c'(x)u(x) + c(x)f(x)u(x)$ et on compare avec (2.5). Ceci implique $c'(x)u(x) = g(x)$ et une simple intégration donne

$$y(x) = C \cdot u(x) + u(x) \int_0^x \frac{g(t)}{u(t)}\, dt. \tag{2.7}$$

Cette relation montre que la solution de (2.5) est la *somme de la solution générale de l'équation homogène et d'une solution particulière de l'équation inhomogène*.

VI.2.4 Équations différentielles d'ordre 2

Une équation différentielle d'ordre 2 est de la forme

$$y'' = f(x, y, y').$$

La résolution analytique d'une telle équation est rarement possible. Il y a pourtant quelques exceptions.

Équations ne dépendant pas de $\mathbf{y}$. Il est naturel de poser $p = y'$ pour que l'équation différentielle $y'' = f(x, y')$ devienne l'équation du premier ordre $p' = f(x, p)$. Remarquons que l'équation (1.2) de la caténaire appartient à ce type.

Équations ne dépendant pas de $\mathbf{x}$. Il s'agit d'une équation différentielle de la forme

$$y'' = f(y, y'). \tag{2.8}$$

L'idée est de considérer y comme variable indépendante et de chercher une fonction $p(y)$ telle que $y' = p(y)$. La règle de la dérivation d'une fonction composée donne

$$y'' = \frac{dp}{dx} = \frac{dp}{dy} \cdot \frac{dy}{dx} = p' \cdot p,$$

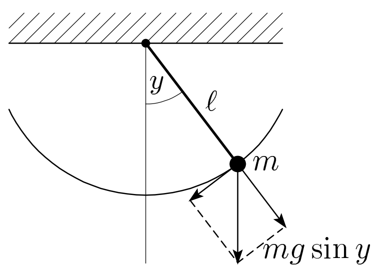
et la formule (2.8) devient une équation du premier ordre

$$p' \cdot p = f(y, p). \tag{2.9}$$

Une fois la solution $p(y)$ de (2.9) trouvée, il reste à intégrer $y' = p(y)$, une équation de type (2.1).

Exemple 2.2 (pendule mathématique) Le mouvement d'un pendule est décrit par l'équation (avec $m = \ell = g = 1$)

$$y'' + \sin y = 0$$



(y désigne l'écart par rapport à la position d'équilibre). Comme cette formule ne dépend pas de t (nous écrivons de nouveau t au lieu de x), nous pouvons utiliser la transformation ci-dessus pour obtenir

$$p \cdot dp = -\sin y \cdot dy \qquad \text{et} \qquad \frac{p^2}{2} = \cos y + C.$$

Si l'on dénote l'amplitude des oscillations par A (le cas où $p = y' = 0$), on obtient $C = -\cos A$ et par conséquent

$$p = \frac{dy}{dt} = \sqrt{2\cos y - 2\cos A} \tag{2.10}$$

qui est à nouveau une équation différentielle pour y. La séparation des variables donne enfin la solution exprimée sous forme implicite à l'aide d'une intégrale elliptique

$$\int_0^y \frac{d\eta}{\sqrt{2\cos\eta - 2\cos A}} = t \tag{2.11}$$

(la constante d'intégration est déterminée par la condition $y = 0$ pour $t = 0$).

Même s'il n'est pas possible d'exprimer l'intégrale de (2.11) en termes de fonctions élémentaires, cette représentation est très utile. Par exemple, si T est la période des oscillations, la déviation maximale A est atteinte pour $t = T/4$. Donc, la période est donnée par (rappelons que $1 - \cos y = 2\sin^2(y/2)$)

$$T = 4\int_0^A \frac{dy}{\sqrt{2\cos y - 2\cos A}} = 2\int_0^A \frac{dy}{\sqrt{\sin^2(A/2) - \sin^2(y/2)}}. \tag{2.12}$$

Avec l'abréviation $\varepsilon := \sin(A/2)$ et en utilisant la substitution $\sin(y/2) = \varepsilon \sin\alpha$, on obtient

$$T = 4\int_0^{\pi/2} \frac{d\alpha}{\sqrt{1 - \varepsilon^2 \sin^2\alpha}} = 4\sum_{k\geq 0} \binom{-1/2}{k} (-1)^k \varepsilon^{2k} \int_0^{\pi/2} \sin^{2k}\alpha \, d\alpha = 2\pi\Big(1 + \frac{A^2}{16} + \frac{11A^4}{3072} + \ldots\Big).$$

Nous voyons que la période T dépend de l'amplitude A ; elle est proche de 2π si A est petit.

VI.3 Équations différentielles linéaires

Soient $a_0(x)$, $a_1(x), \ldots, a_{n-1}(x)$ des fonctions données. On appelle

$$y^{(n)} + a_{n-1}(x)y^{(n-1)} + \ldots + a_1(x)y' + a_0(x)y = 0 \tag{3.1}$$

une *équation différentielle linéaire homogène* d'ordre n et

$$y^{(n)} + a_{n-1}(x)y^{(n-1)} + \ldots + a_1(x)y' + a_0(x)y = f(x) \tag{3.2}$$

une *équation différentielle linéaire inhomogène* (si $f(x) \not\equiv 0$). Pour le membre de gauche de ces équations, introduisons le symbole

$$\mathcal{L}(y) := y^{(n)} + a_{n-1}(x)y^{(n-1)} + \ldots + a_0(x)y. \tag{3.3}$$

Les équations (3.1) et (3.2) s'écrivent alors

$$\mathcal{L}(y) = 0 \quad \text{respectivement} \quad \mathcal{L}(y) = f. \tag{3.4}$$

L'*opérateur différentiel* $\mathcal{L}$ opère sur des fonctions $y(x)$ et le résultat $\mathcal{L}(y)$ est de nouveau une fonction, donnée par (3.3). Il a une propriété importante, la *linéarité*, i.e.

$$\mathcal{L}(c_1y_1 + c_2y_2) = c_1\mathcal{L}(y_1) + c_2\mathcal{L}(y_2). \tag{3.5}$$

Le résultat suivant est une conséquence immédiate de cette linéarité.

Lemme 3.1 *Soient $y_1(x),\ y_2(x), \ldots, y_n(x)$ n solutions de l'équation homogène (3.1). Alors, pour des constantes arbitraires $c_1, \ldots, c_n$, la fonction*

$$c_1y_1(x) + c_2y_2(x) + \ldots + c_ny_n(x) \tag{3.6}$$

est aussi une solution de cette équation. □

Remarque. La solution d'une équation d'ordre 1 comporte une constante (voir le paragraphe VI.2) ; la solution d'une équation d'ordre 2 comporte deux constantes arbitraires (voir par exemple la formule (1.3)). Par analogie, on peut supposer (Euler) que la solution d'une équation d'ordre n comporte n constantes et que (3.6) est *la solution générale* de (3.1) (si $y_1(x), \dots, y_n(x)$ sont linéairement indépendantes).[3]

Lemme 3.2 *Nous avons que*

$$\left\{ \begin{array}{c} \textit{solution générale de} \\ \textit{l'équat. inhomogène (3.2)} \end{array} \right\} = \left\{ \begin{array}{c} \textit{solution générale de} \\ \textit{l'équat. homogène (3.1)} \end{array} \right\} + \left\{ \begin{array}{c} \textit{solution particulière de} \\ \textit{l'équat. inhomogène (3.2)} \end{array} \right\}$$

Démonstration. Soit $\widetilde{y}$ une solution particulière de (3.2), i.e. $\mathcal{L}(\widetilde{y}) = f$. Pour une solution arbitraire y de (3.1) (i.e. $\mathcal{L}(y) = 0$), nous avons $\mathcal{L}(y + \widetilde{y}) = f$ par (3.5) et $y + \widetilde{y}$ est une solution de (3.2).

D'autre part, si $\widehat{y}$ est une autre solution de (3.2) (i.e. $\mathcal{L}(\widehat{y}) = f$) alors, de nouveau par (3.5), nous avons $\mathcal{L}(\widehat{y} - \widetilde{y}) = 0$ et $\widehat{y} = \widetilde{y} + (\widehat{y} - \widetilde{y})$ est la somme de $\widetilde{y}$ et d'une solution de l'équation homogène (3.1). □

Conclusion. Il faut trouver, pour résoudre entièrement les équations (3.1) et (3.2) :
- n solutions linéairement indépendantes de (3.1),
- *une* solution de (3.2).

VI.3.1 Équations homogènes à coefficients constants

La résolution complète de l'équation (3.1) en formules analytiques est rarement possible. Il y a pourtant des exceptions, dont la plus importante se présente lorsque les coefficients $a_i(x)$ ne dépendent pas de x, i.e.

$$y^{(n)} + a_{n-1}y^{(n-1)} + \ldots + a_1y' + a_0y = 0. \tag{3.7}$$

L'idée essentielle pour la résolution de (3.7) consiste à rechercher des solutions de la forme

$$y(x) = e^{\lambda x}, \tag{3.8}$$

où λ est une constante à déterminer. En insérant les dérivées

$$y'(x) = \lambda e^{\lambda x}, \qquad y''(x) = \lambda^2 e^{\lambda x}, \quad \ldots \quad , \ y^{(n)}(x) = \lambda^n e^{\lambda x},$$

dans l'équation (3.7), on obtient

$$(\lambda^n + a_{n-1}\lambda^{n-1} + \ldots + a_1\lambda + a_0)e^{\lambda x} = 0. \tag{3.9}$$

La fonction (3.8) est donc une solution de (3.7), si et seulement si λ est une solution de l'*équation caractéristique*

$$\chi(\lambda) = 0, \qquad \chi(\lambda) := \lambda^n + a_{n-1}\lambda^{n-1} + \ldots + a_1\lambda + a_0. \tag{3.10}$$

Racines distinctes. Si l'équation (3.10) admet n racines distinctes $\lambda_1, \ldots, \lambda_n$, alors $e^{\lambda_1 x}, \ldots, e^{\lambda_n x}$ sont n solutions linéairement indépendantes de (3.7). La solution générale est alors donnée par

$$y(x) = c_1e^{\lambda_1 x} + c_2e^{\lambda_2 x} + \ldots + c_ne^{\lambda_n x}. \tag{3.11}$$

[3] On dit que les fonctions $y_1(x), \ldots, y_n(x)$ sont *linéairement indépendantes* si la combinaison linéaire (3.6) est identiquement nulle seulement si tous les c_i sont nuls. Par exemple, $1, x, x^2, x^3$ sont des fonctions linéairement indépendantes.

Racines multiples. Examinons pour commencer le cas simple de l'équation différentielle

$$y^{(n)} = 0, \tag{3.12}$$

dont l'équation caractéristique $\lambda^n = 0$ possède une racine, de multiplicité n. La solution générale de (3.12) est manifestement $c_1 + c_2 x + c_3 x^2 + \ldots + c_n x^{n-1}$, un polynôme de degré $n-1$.

Étudions ensuite l'équation

$$y''' - 3ay'' + 3a^2 y' - a^3 y = 0, \tag{3.13}$$

dont l'équation caractéristique $(\lambda - a)^3 = 0$ possède la racine a de multiplicité 3. Nous pouvons alors introduire une nouvelle fonction $u(x)$ en écrivant

$$y(x) = e^{ax} \cdot u(x). \tag{3.14}$$

Si nous dérivons $u(x) = e^{-ax} \cdot y(x)$ trois fois, nous obtenons pour $u(x)$ l'équation $u''' = 0$. La solution générale de (3.13) est donc donnée par

$$y(x) = e^{ax} \cdot (c_1 + c_2 x + c_3 x^2). \tag{3.15}$$

Pour le cas général on a le résultat suivant.

Théorème 3.3 *Si le polynôme caractéristique (3.10) se factorise sous la forme*

$$\chi(\lambda) = (\lambda - \lambda_1)^{m_1} (\lambda - \lambda_2)^{m_2} \cdot \ldots \cdot (\lambda - \lambda_k)^{m_k}$$

(avec des λ_i distincts), alors la solution générale de (3.7) est donnée par

$$y(x) = p_1(x) e^{\lambda_1 x} + p_2(x) e^{\lambda_2 x} + \ldots + p_k(x) e^{\lambda_k x}, \tag{3.16}$$

où les $p_i(x)$ sont des polynômes arbitraires de degré $m_i - 1$ (cette solution contient exactement $\sum_{i=1}^{k} m_i = n$ constantes libres).

Démonstration. Nous illustrons la démonstration dans le cas de deux racines multiples $\chi(\lambda) = (\lambda - a)^3 (\lambda - b)^2$. L'équation différentielle peut être écrite sous la forme

$$(y''' - 3ay'' + 3a^2 y' - a^3 y)'' - 2b(y''' - 3ay'' + 3a^2 y' - a^3 y)' + b^2 (y''' - 3ay'' + 3a^2 y' - a^3 y) = 0.$$

Chaque solution de (3.13) est donc aussi solution de cette équation différentielle. Écrit sous la forme

$$(y'' - 2by' + b^2 y)''' - 3a(y'' - 2by' + b^2 y)'' + 3a^2 (y'' - 2by' + b^2 y)' - a^3 (y'' - 2by' + b^2 y) = 0,$$

on voit que $e^{bx}(d_1 + d_2 x)$ est également une solution. L'affirmation est donc une conséquence de la linéarité. □

Éviter l'arithmétique complexe. Le résultat du théorème 3.3 est également valable pour des λ_i complexes. Cependant, si les coefficients a_i de l'équation (3.7) sont réels, nous cherchons avant tout à obtenir des solutions réelles. Le fait que les racines complexes des polynômes réels apparaissent par paires conjuguées nous permet de simplifier (3.16). Soient $\lambda_1 = \alpha + i\beta$ et $\lambda_2 = \alpha - i\beta$ deux racines complexes conjuguées. Les termes de la solution (3.16) qui correspondent à ces racines sont alors donnés par un polynôme multiplié par

$$e^{\alpha x} (c_1 e^{i\beta x} + c_2 e^{-i\beta x}). \tag{3.17}$$

Avec la formule d'Euler $e^{i\beta x} = \cos \beta x + i \sin \beta x$, cette expression s'écrit

$$e^{\alpha x} (d_1 \cos \beta x + d_2 \sin \beta x) = C e^{\alpha x} \Big(\sin \varphi \cos \beta x + \cos \varphi \sin \beta x \Big) = C e^{\alpha x} \sin(\beta x + \varphi),$$

où $d_1 = c_1 + c_2$ et $d_2 = i(c_1 - c_2)$ sont de nouvelles constantes. En utilisant $d_2 + i d_1 = C e^{i\varphi} = C \cos \varphi + i C \sin \varphi$, l'expression devient encore plus simple. Voir la figure VI.2 pour un dessin de cette solution.

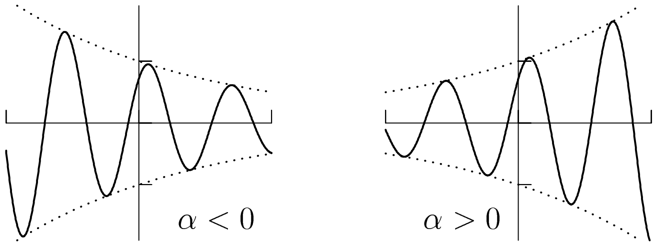


Figure VI.2: Oscillations stable et instable

VI.3.2 Équations linéaires inhomogènes

Le problème consiste à trouver *une* solution particulière de $\mathcal{L}(y) = f$, i.e. de

$$y^{(n)} + a_{n-1}y^{(n-1)} + \ldots + a_1 y' + a_0 y = f(x). \tag{3.18}$$

Le résultat suivant est une conséquence immédiate de la linéarité de (3.5).

Lemme 3.4 (Principe de superposition) *Si $y_1(x)$ et $y_2(x)$ sont des solutions de $\mathcal{L}(y_1) = f_1$ et $\mathcal{L}(y_2) = f_2$, alors $c_1y_1(x) + c_2y_2(x)$ est une solution de $\mathcal{L}(y) = c_1f_1 + c_2f_2$.* □

Il y a deux approches pour déterminer une solution particulière de (3.18):

- La *méthode de la variation des constantes*. On a vu cette technique pour des équations de premier ordre (paragraphe VI.2.3) et nous allons la voir pour des systèmes d'équations différentielles linéaires dans le paragraphe VI.6.
- La *méthode rapide*. Elle est très efficace, mais elle est applicable seulement si le second membre $f(x)$ dans (3.18) est une somme de termes simples. Par le lemme 3.4 ces termes peuvent être traités séparément.

La méthode rapide. C'est une approche qui est possible si $f(x)$ est une combinaison linéaire de x^j, e^{ax}, $e^{\alpha x}\sin(\omega x), \ldots$, plus précisément si $f(x)$ est elle-même solution d'une équation linéaire homogène à coefficients constants. On essaie de trouver une solution avec la même structure.

Exemple 3.5 Considérons le cas où $f(x)$ est un polynôme de degré 2, par exemple

$$y''' + 5y'' + 2y' + y = 2x^2 + x. \tag{3.19}$$

Cherchons une solution de la forme

$$y(x) = a + bx + cx^2. \tag{3.20}$$

En insérant les dérivées de (3.20) dans (3.19), il vient

$$cx^2 + (b + 4c)x + (a + 2b + 10c) = 2x^2 + x.$$

En comparant les coefficients des mêmes puissances de x, on a $c = 2$, $b = -7$ et $a = -6$. Une solution particulière de (3.19) est donc

$$y(x) = 2x^2 - 7x - 6.$$

Exemple 3.6 Supposons que $f(x)$ soit une fonction sinus,

$$y'' - y' + y = \sin 2x. \tag{3.21}$$

Il ne suffit pas de prendre $y(x) = a \cdot \sin 2x$, puisque y' contient $\cos 2x$. Posons donc

$$y(x) = a \cdot \sin 2x + b \cdot \cos 2x, \tag{3.22}$$

calculons les dérivées et insérons-les dans (3.21). Nous obtenons la condition

$$(a + 2b - 4a) \sin 2x + (b - 2a - 4b) \cos 2x = \sin 2x.$$

Nous en déduisons le système linéaire $-3a+2b = 1$, $-2a-3b = 0$, dont la solution est $a = -3/13$, $b = 2/13$. Par conséquent, une solution particulière de (3.21) est

$$y(x) = -\frac{3}{13} \sin 2x + \frac{2}{13} \cos 2x. \tag{3.23}$$

Une autre possibilité pour résoudre (3.21) consiste à considérer l'équation

$$y'' - y' + y = e^{2ix} \tag{3.24}$$

et à en chercher une solution de la forme $y(x) = Ae^{2ix}$. En insérant les dérivées de Ae^{2ix} dans (3.24), nous obtenons $-4A - 2iA + A = 1$ et $A = (-3 + 2i)/13$. Une solution de (3.24) est donc

$$y(x) = \frac{-3 + 2i}{13} e^{2ix}. \tag{3.25}$$

Comme (3.21) est la partie imaginaire de (3.24), nous obtenons la solution de (3.21) en prenant la partie imaginaire de (3.25).

Justification de cette approche. Par hypothèse, $f(x)$ satisfait $\mathcal{L}_1(f) = 0$, où $\mathcal{L}_1$ est un opérateur différentiel à coefficients constants. En appliquant cet opérateur à l'équation (3.18), i.e. à $\mathcal{L}(y) = f$, nous obtenons $(\mathcal{L}_1\mathcal{L})(y) = 0$, et la solution de (3.18) est une solution de l'équation différentielle linéaire homogène $(\mathcal{L}_1\mathcal{L})(y) = 0$. La solution générale de cette équation est donnée par le théorème 3.3.

Cas de résonance. Considérons par exemple l'équation

$$y'' + y = \sin x. \tag{3.26}$$

Il ne suffit pas de poser ici $y(x) = a \sin x + b \cos x$, car cette fonction est une *solution de l'équation homogène*. La justification ci-dessus (voir aussi figure VI.3) nous suggère de tenter

$$y(x) = ax \sin x + bx \cos x. \tag{3.27}$$

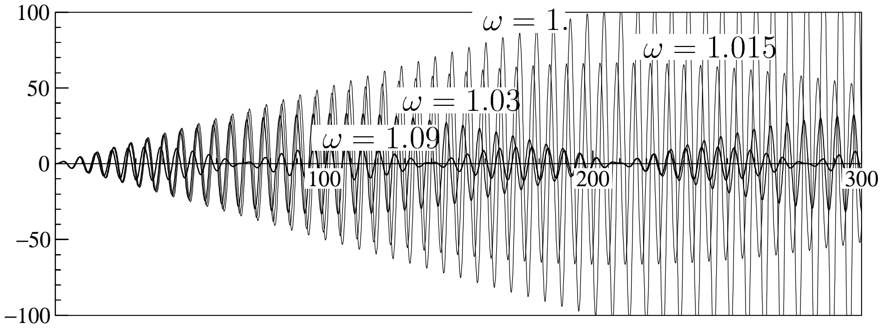


Figure VI.3: Solution de $y'' + y = \sin \omega x$, $y(0) = 0$, $y'(0) = 1$, $\omega = 1.09, 1.03, 1.015, 1.$

La méthode usuelle (insérer les dérivées de (3.27) dans (3.26)) livre $2a\cos x - 2b\sin x = \sin x$, et donc $a = 0$ et $b = -1/2$. Ainsi, une solution particulière de (3.26) est donnée par

$$y(x) = -\frac{1}{2}\, x\cos x. \tag{3.28}$$

Elle explose pour $x \to \infty$ (cf. figure VI.3).

VI.4 Systèmes d'équations différentielles – exemples

Dans les premiers paragraphes, nous avons considéré des équations différentielles pour une fonction scalaire. A partir de maintenant nous étudions des problèmes

$$y' = f(x, y) \tag{4.1}$$

où $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$. On cherche une fonction vectorielle $y : I \to \mathbb{R}^n$ définie sur un intervalle I et satisfaisant $y'(x) = f(x, y(x))$ pour tout $x \in I$.

Remarquons qu'une équation différentielle d'ordre supérieur, par exemple

$$z''' = g(x, z, z, z''),$$

peut être transformée en un système (4.1) d'ordre un. En posant $y_1 = z$, $y_2 = z'$ et $y_3 = z''$, on obtient

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}' = \begin{pmatrix} y_2 \\ y_3 \\ g(x, y_1, y_2, y_3) \end{pmatrix}$$

qui est sous la forme souhaitée.

VI.4.1 Le problème de Lotka–Volterra

Commençons par un problème de la biologie mathématique. Le développement de deux populations (par exemple, $y_1(t)$ pour le nombre de lynx et $y_2(t)$ pour le nombre de lièvres) peut être modélisé par les équations de Lotka-Volterra

$$y_1' = y_1(\alpha y_2 - \beta), \qquad y_2' = y_2(\gamma - \delta y_1), \tag{4.2}$$

où $\alpha, \beta, \gamma, \delta$ sont des constantes positives. Ceci implique que la population y_1 croît si y_2 est plus grand qu'une certaine valeur de seuil (β/α) alors que sinon elle décroît. Pour l'autre population, c'est le contraire.

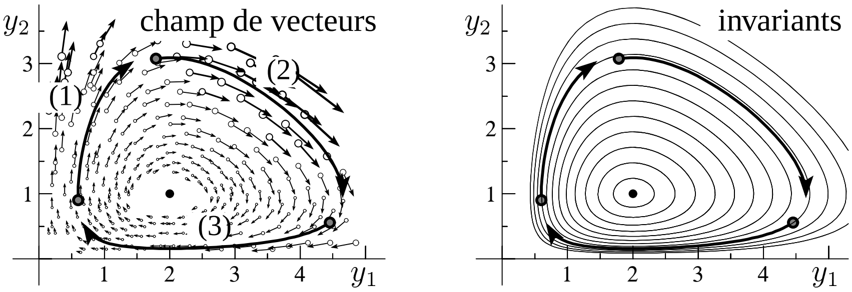


Figure VI.4: Champ de vecteurs (gauche) et invariants (droite) pour le problème (4.2) de Lotka-Volterra avec $\alpha = \beta = \delta = 1$ et $\gamma = 2$.

Cette équation différentielle est sous la forme $y' = f(y)$ où $f : \mathbb{R}^2 \to \mathbb{R}^2$. Elle est représentée graphiquement dans la figure VI.4 (dessin de gauche). Pour un point $y = (y_1, y_2)^t$, le vecteur $f(y)$ est dessiné comme une flèche attachée au point y. L'équation différentielle exprime le fait que la solution est obligée de suivre ces flèches.

On voit sur la figure VI.4 que la solution du système (4.2) traverse trois étapes: (1) si la population de lynx est petite, celle des lièvres croît; (2) dès qu'il y a suffisamment de lièvres, la population de lynx croît entraînant une diminution de celle des lièvres; (3) la population des lynx décroît à cause du manque de lièvres. Ces trois étapes se répètent cycliquement.

Invariant du problème de Lotka-Volterra. Pour étudier les solutions, divisons les deux équations de (4.2) et considérons y_1 comme fonction de y_2 (ou y_2 comme fonction de y_1). On obtient ainsi (par séparation de variables)

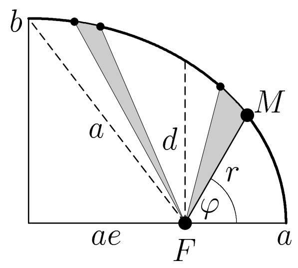
$$\frac{dy_1}{dy_2} = \frac{y_1(\alpha y_2 - \beta)}{y_2(\gamma - \delta y_1)} \qquad \text{ou} \qquad \frac{(\gamma - \delta y_1)}{y_1}\, dy_1 = \frac{(\alpha y_2 - \beta)}{y_2}\, dy_2.$$

Une intégration de la dernière équation donne

$$\gamma \ln y_1 - \delta\, y_1 = \alpha\, y_2 - \beta \ln y_2 + Const. \tag{4.3}$$

Dans la figure VI.4 (droite) sont dessinées des courbes de niveau de la fonction (4.3). Chaque solution $(y_1(t), y_2(t))^t$ reste sur une courbe de niveau pour tout $t \in \mathbb{R}$. Ceci suggère que les solutions de (4.2) sont exactement périodiques.

VI.4.2 Le problème de Kepler



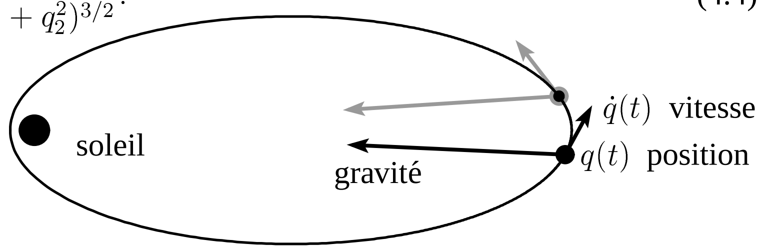
Une des plus grandes découvertes de l'histoire de la science, due à J. Kepler (1609) et basée sur des mesures précises de positions de la planète mars faites par lui-même et par Tycho Brahe, est que les orbites des planètes sont elliptiques avec le soleil dans un des foyers (première loi de Kepler):

$$r = \frac{d}{1 + e\cos\varphi},$$

(dans le dessin, $a =$ le grand axe, $e =$ l'excentricité, $b = a\sqrt{1-e^2}$, $d = b\sqrt{1-e^2} = a(1-e^2)$). La deuxième loi de Kepler ($r^2\dot\varphi = Const$) dit que le segment qui joint une planète au soleil balaie des aires égales en des intervalles de temps égaux.

Newton (*Principia* 1687) a réussi à expliquer ce mouvement des planètes par sa loi de gravité (force est proportionelle à $1/r^2$) et par la relation entre forces et acceleration (la "Lex II" de la *Principia*). Si l'on choisit le soleil comme centre du système des coordonnées, la planète restera dans un plan et nous pouvons décrire sa position par deux coordonnées (q_1, q_2). Avec des normalisations convenables, ceci donne le système d'équations différentielles d'ordre deux

$$\ddot q_1 = -\frac{q_1}{(q_1^2 + q_2^2)^{3/2}}, \qquad \ddot q_2 = -\frac{q_2}{(q_1^2 + q_2^2)^{3/2}}. \tag{4.4}$$



On peut introduire des vitesses

$$v_1 = \dot q_1, \qquad v_2 = \dot q_2$$

pour obtenir un système d'ordre un.

Invariants du problème de Kepler. Cette équation différentielle possède plusieurs invariants qui permettent une solution analytique. On vérifie par différentiation que les expressions

$$\frac{1}{2}(v_1^2+v_2^2)-\frac{1}{\sqrt{q_1^2+q_2^2}}=H_0 \qquad \text{(énergie totale)}$$

$$q_1v_2-q_2v_1=L_0 \qquad \text{(moment cinétique)}$$

$$\begin{pmatrix} v_2 \\ -v_1 \end{pmatrix} L_0-\frac{1}{\sqrt{q_1^2+q_2^2}}\begin{pmatrix} q_1 \\ q_2 \end{pmatrix}=\begin{pmatrix} A_{10} \\ A_{20} \end{pmatrix} \qquad \text{(vecteur de Runge-Lenz-Pauli)}$$

sont constantes le long des solutions de (4.4). Par exemple, la dérivée du moment cinétique donne

$$\frac{d}{dt}\Big(q_1v_2-q_2v_1\Big)=\dot q_1v_2+q_1\dot v_2-\dot q_2v_1-q_2\dot v_1=v_1v_2-\frac{q_1q_2}{\sqrt{q_1^2+q_2^2}}-v_2v_1-\frac{q_2q_1}{\sqrt{q_1^2+q_2^2}}=0.$$

Deuxième loi de Kepler. Ecrit en coordonnées polaires $q_1=r\cos\varphi$, $q_2=r\sin\varphi$, $\dot q_1=\dot r\cos\varphi-r\dot\varphi\sin\varphi$, $\dot q_2=\dot r\sin\varphi+r\dot\varphi\cos\varphi$, la formule pour le moment cinétique donne

$$L_0=q_1\dot q_2-q_2\dot q_1=\ldots=r^2\dot\varphi$$

ce qui démontre la deuxième loi de Kepler.

Orbite elliptique. La formule pour le vecteur de Runge-Lenz-Pauli nous permet d'exprimer L_0v_1 et L_0v_2 en termes de q_1 et q_2. Ces expressions, insérées dans $L_0(q_1v_2-q_2v_1)=L_0^2$ donnent

$$L_0^2=q_1\Big(A_{10}+\frac{q_1}{\sqrt{q_1^2+q_2^2}}\Big)+q_2\Big(A_{20}+\frac{q_2}{\sqrt{q_1^2+q_2^2}}\Big)=q_1A_{10}+q_2A_{20}+\sqrt{q_1^2+q_2^2}.$$

En sortant la racine et en prenant son carré, on obtient une équation quadratique pour (q_1,q_2), ce qui montre que l'orbite est bien une conique.

VI.4.3 Le système solaire (problème à N corps)

Le problème de Kepler traîte le mouvement d'une planète autour du soleil en tenant compte uniquement de la force d'attraction entre le soleil et la planète. En utilisant les lois de Newton, on peut formuler sans difficulté les équations décrivant le mouvement en présence des plusieurs planètes. Il faut simplement additionner toutes les forces agissant sur un corps (voir la figure VI.5).

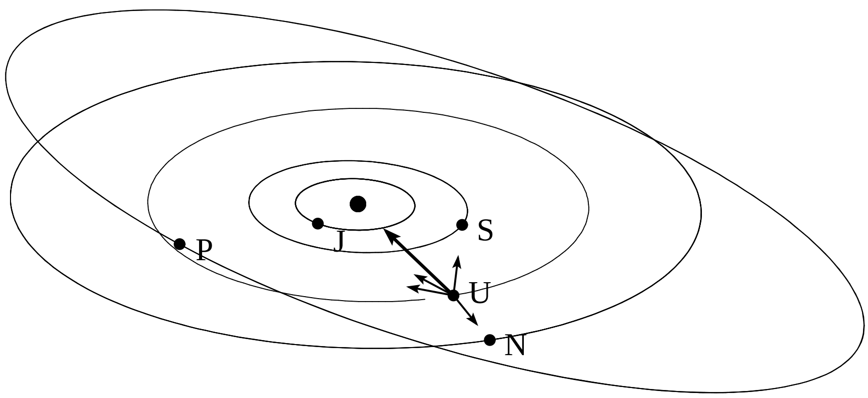


Figure VI.5: Le système solaire avec les cinq planètes extérieures

En notant par $q_i \in \mathbb{R}^3$ la position de la $i$ème planète ($q_0 \in \mathbb{R}^3$ pour le soleil) et par m_i sa masse, la loi d'attraction et la deuxième loi de Newton donnent l'équation differentielle

$$m_i\, \ddot{q}_i = -\sum_{i=1}^{5}\sum_{j=0}^{i-1} G\, m_i\, m_j\, \frac{q_i - q_j}{\|q_i - q_j\|^3}, \qquad i = 0, \ldots, 5. \tag{4.5}$$

On peut de nouveau trouver quelques invariants (énergie totale, moment cinétique), mais ils ne suffisent pas pour pouvoir résoudre analytiquement ce problème. Pour prédire des eclipses du soleil ou la stabilité du système solaire sur des millions d'années, on est donc obligé d'utiliser des méthodes numériques (cours de deuxième année). Dans le paragraphe suivant nous allons étudier l'existence et l'unicité de la solution d'une telle équation différentielle.

VI.4.4 Réactions chimiques

Considérons un mélange de trois substances chimiques qui réagissent selon les formules suivantes:

$$\begin{array}{ccll} A & \xrightarrow{0.04} & B & \text{(lente)} \\ B + B & \xrightarrow{3\cdot 10^7} & C + B & \text{(très rapide)} \\ B + C & \xrightarrow{10^4} & A + C & \text{(rapide)} \end{array} \tag{4.6}$$

Notons les concentrations de A, B et C par y_1, y_2 et y_3. Une loi de la chimie ("differentiation law") nous donne l'équation différentielle

$$\begin{array}{lllll} \text{A:} & y_1' = & -0.04 y_1 + 10^4 y_2 y_3 & & y_1(0) = 1 \\ \text{B:} & y_2' = & 0.04 y_1 - 10^4 y_2 y_3 & -3\cdot 10^7 y_2^2 & y_2(0) = 0 \\ \text{C:} & y_3' = & & 3\cdot 10^7 y_2^2 & y_3(0) = 0. \end{array} \tag{4.7}$$

Elle est obtenue comme suit: pour chaque réaction on considère le produit des concentrations de substances apparaissant à gauche dans la formule chimique, on le multiplie par la constante décrivant la vitesse de réaction, et on ajoute ce produit avec le facteur $k - \ell$ à l'équation pour la substance D, si D apparaît k fois à droite et ℓ fois à gauche de la formule chimique.

On a peu d'espoir de résoudre analytiquement ce problème et on est restreint à étudier des propriétés théoriques (existence, unicité de la solution, stabilité, ...). Pour obtenir des résultats quantitatifs, on est obligé d'utiliser des méthodes numériques.

VI.5 Existence et unicité du problème de Cauchy

Considérons le problème de Cauchy

$$y' = f(x, y), \qquad y(x_0) = y_0 \tag{5.1}$$

où $f : U \to \mathbb{R}^n$ (avec $U \subset \mathbb{R} \times \mathbb{R}^n$ ouvert) est une fonction continue. En intégrant l'équation différentielle entre x_0 et x on obtient *l'équation intégrale*

$$y(x) = y_0 + \int_{x_0}^{x} f(t, y(t))\, dt. \tag{5.2}$$

Chaque solution de (5.1) est donc solution de (5.2). Le contraire est vrai aussi. Si une fonction continue $y(x)$ vérifie (5.2) sur un intervalle I, alors elle est automatiquement continûment différentiable et elle vérifie (5.1).

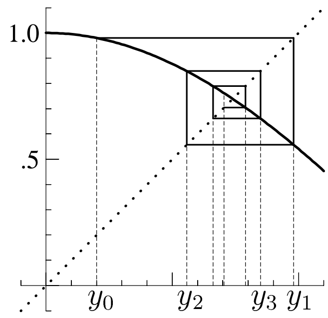
L'équation (5.2) est sous la forme $y = G(y)$ où pour une fonction $y(x)$ donnée, $G(y)$ est la fonction définie par le membre de droite de (5.2). Ceci suggère d'utiliser la méthode des approximations successives.

Rappel: la méthode des approximations successives. Pour résoudre un problème $y = g(y)$ avec $g : \mathbb{R} \to \mathbb{R}$, cette méthode est définie par:

- on choisit y_0 arbitrairement,
- on applique l'itération $y_{k+1} = g(y_k)$.

Si cette itération converge, on est sûr d'avoir trouvé une solution (si g est continue). La solution n'est pas nécessairement unique.

Prenons, par exemple, la fonction $g(y) = \cos y$. Par le théorème des accroissements finis, on a $|g(y) - g(z)| \le \gamma |y - z|$ avec $\gamma = \sin 1 < 1$ (pour $y, z \in [0, 1]$). Ceci permet de démontrer que $\{y_k\}$ est une suite de Cauchy et donc qu'elle converge (voir le petit dessin).



Itération de Picard-Lindelöf L'équation (5.2) peut être considérée comme un problème à point fixe. L'idée est d'appliquer la méthode des approximations successives. Elle s'écrit sous la forme

$$\begin{aligned} y_0(x) &= y_0 \qquad \text{(ou une fonction arbitraire)} \\ y_{k+1}(x) &= y_0 + \int_{x_0}^{x} f(t, y_k(t))\, dt. \end{aligned} \tag{5.3}$$

Exemple 5.1 Considérons le problème

$$y' = -y^2, \qquad y(0) = 1$$

avec comme solution exacte $y(x) = 1/(1 + x)$. Les premières approximations obtenues par l'itération de Picard-Lindelöf sont $y_0(x) = 1$, $y_1(x) = 1 - x$ et $y_2(x) = 1 - x + x^2 - x^3/3$ (voir la figure VI.6). On observe une convergence rapide vers la solution exacte dans l'intervalle $[0, 3.75]$. Pour x trop grand, l'itération diverge.

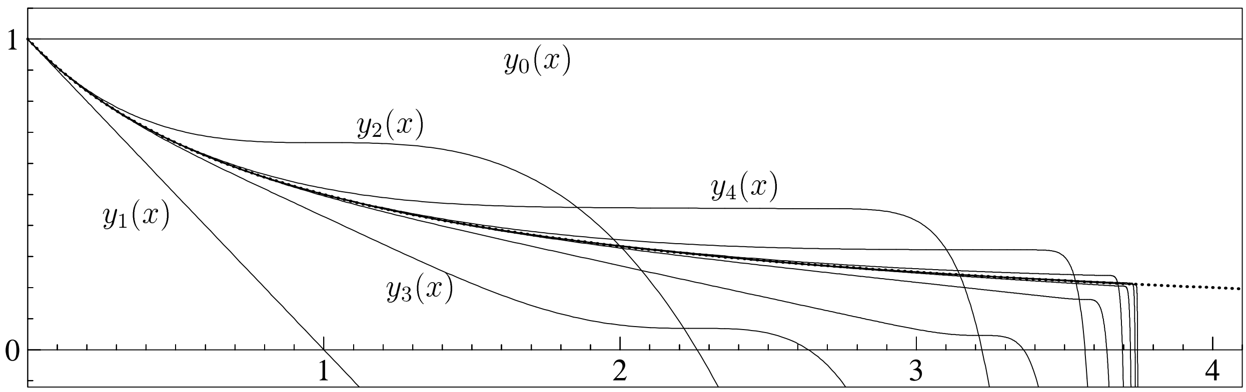


Figure VI.6: Itération de Picard-Lindelöf pour le problème de l'Exemple 5.1

Lemme 5.2 *Soit* $A = \{(x, y) \in \mathbb{R} \times \mathbb{R}^n\,;\ |x - x_0| \le a,\ \|y - y_0\| \le b\}$, $f : A \to \mathbb{R}^n$ *une fonction continue et* $M = \max_{(x,y)\in A} \|f(x, y)\|$. *Pour* $\alpha := \min(a, b/M)$, *les fonctions* $y_k(x)$ *de l'itération de Picard-Lindelöf sont bien définie pour* $|x - x_0| \le \alpha$ *et elles satisfont*

$$\|y_k(x) - y_0\| \le b \quad \textit{pour} \quad |x - x_0| \le \alpha.$$

Démonstration. L'affirmation est obtenue par récurrence sur k comme suit

$$\|y_{k+1}(x) - y_0\| = \Big\| \int_{x_0}^{x} f(t, y_k(t))\, dt \Big\| \le \int_{x_0}^{x} \|f(t, y_k(t))\|\, dt \le M|x - x_0| \le M\alpha \le b. \qquad \square$$

On dit qu'une fonction $f : A \to \mathbb{R}^n$ (avec A comme dans le lemme précédent) satisfait une *condition de Lipschitz* si

$$\|f(x,y) - f(x,z)\| \le L\,\|y-z\| \qquad \text{pour } \ (x,y),(x,z) \in A. \tag{5.4}$$

La constante L s'appelle *constante de Lipschitz*. Remarquons que la condition (5.4) n'est pas une conséquence de la continuité de $f(x,y)$. Par exemple, la fonction $\sqrt{|y|}$ est continue sans vérifier (5.4). D'autre part, chaque fonction qui est continûment différentiable vérifie (5.4). Ceci est une conséquence du théorème des accroissements finis (théorème III.5.1).

Théorème 5.3 (existence et unicité du problème de Cauchy) *Considérons l'ensemble* $A = \{(x,y) \in \mathbb{R}\times\mathbb{R}^n\,;\, |x-x_0| \le a, \|y-y_0\| \le b\}$ *et supposons que* $f : A \to \mathbb{R}^n$

- *soit continue,*
- *satisfasse une condition de Lipschitz.*

Alors, le problème de Cauchy $y' = f(x,y),\ y(x_0) = y_0$ *possède une solution unique sur* $I = [x_0-\alpha, x_0+\alpha]$*, où* $\alpha = \min(a, b/M)$ *et* $M = \max_{(x,y)\in A} \|f(x,y)\|$.

Démonstration. Existence. L'idée est de démontrer que l'itération de Picard-Lindelöf converge uniformément sur I vers une solution du problème de Cauchy. Dans une première partie nous allons démontrer que

$$\|y_{k+1}(x) - y_k(x)\| \le ML^k\,\frac{|x-x_0|^{k+1}}{(k+1)!} \qquad \text{pour } \ |x-x_0| \le \alpha. \tag{5.5}$$

Pour $k = 0$ avec pour $y_0(x) = y_0$, cette estimation suit de $\|\int_{x_0}^x f(t,y(t))\,dt\| \le M\,|x-x_0|$. Supposons qu'elle soit vraie pour $k-1$. Alors, on a pour $x_0 \le x \le x_0+\alpha$ que

$$\begin{aligned}\|y_{k+1}(x) - y_k(x)\| &\le \int_{x_0}^x \|f(t,y_k(t)) - f(t,y_{k-1}(t))\|\,dt \le L\int_{x_0}^x \|y_k(t) - y_{k-1}(t)\|\,dt\\ &\le ML^k\int_{x_0}^x \frac{|t-x_0|^k}{k!}\,dt = ML^k\,\frac{|x-x_0|^{k+1}}{(k+1)!}\end{aligned}$$

(la démonstration pour $x_0 - \alpha \le x \le x_0$ est analogue).

De l'estimation (5.5) nous déduisons que $\{y_k(x)\}$ est une suite de Cauchy qui converge uniformément. En effet,

$$\begin{aligned}\|y_{k+m}(x) - y_k(x)\| &\le \|y_{k+m}(x) - y_{k+m-1}(x)\| + \ldots + \|y_{k+1}(x) - y_k(x)\|\\ &\le \frac{M}{L}\Big(\frac{(L|x-x_0|)^{k+m}}{(k+m)!} + \ldots + \frac{(L|x-x_0|)^{k+1}}{(k+1)!}\Big) \le \frac{M}{L}\sum_{j\ge k+1}\frac{(L\alpha)^j}{j!},\end{aligned}$$

ce qui est le reste d'une série convergente. Donc, cette expression est plus petite que ε si k est suffisamment grand. Comme la convergence est uniforme, la suite $\{y_k(x)\}$ converge vers une fonction continue $y : I \to \mathbb{R}^n$ (théorème I.5.1).

Pour démontrer que cette fonction $y(x)$ est une solution du problème de Cauchy, nous passons à la limite $k \to \infty$ dans (5.3). Comme $\{y_k(x)\}$ converge uniformément et $f(x,y)$ satisfait la condition de Lipschitz (5.4), la suite $\{f(x,y_k(x))\}$ converge uniformément vers $f(x,y(x))$. On peut donc échanger la limite avec l'intégration dans (5.3) et on voit que $y(x)$ est solution de l'équation intégrale (5.2).

Unicité. Supposons que $y(x)$ et $z(x)$ soient deux solutions sur I. Ceci implique que $y(x) - z(x) = \int_{x_0}^{x}(f(t,y(t)) - f(t,z(t)))\,dt$. On en déduit que $\|y(x) - z(x)\| \le 2M|x - x_0|$. Comme dans la première partie de la démonstration nous trouvons par récurrence que pour tout $k \ge 0$

$$\|y(x) - z(x)\| \le 2ML^k \frac{|x - x_0|^{k+1}}{(k+1)!} \le \frac{2M}{L}\,\frac{(L\alpha)^{k+1}}{(k+1)!}.$$

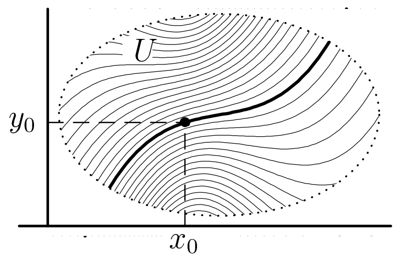
La limite $k \to \infty$ montre que $\|y(x) - z(x)\| = 0$ pour tout $x \in I$. □

VI.5.1 Prolongement des solutions et existence globale

Le théorème 5.3 garantit *l'existence locale* d'une solution du problème de Cauchy. Si la fonction $f(x,y)$ est continûment différentiable dans un voisinage de (x_0, y_0), le problème de Cauchy $y' = f(x,y)$, $y(x_0) = y_0$ possède alors une solution unique sur $[x_0 - \alpha, x_0 + \alpha]$ avec un $\alpha > 0$. Considérons maintenant le point $x_1 = x_0 + \alpha$, $y_1 = y(x_0 + \alpha)$ sur cette solution. On peut réappliquer le théorème 5.3 au problème de Cauchy $y' = f(x,y)$, $y(x_1) = y_1$ et ainsi prolonger cette solution. Jusqu'où peut on prolonger la solution de cette manière?

Théorème 5.4 *Supposons que la fonction $f : U \to \mathbb{R}^n$ (U un ouvert de $\mathbb{R} \times \mathbb{R}^n$) soit continûment différentiable sur U. Alors,*

- *chaque solution de $y' = f(x,y)$ peut être prolongée jusqu'au bord de U,*
- *deux solutions de $y' = f(x,y)$ ne s'intersectent jamais.*



La deuxième affirmation est une conséquence immédiate du théorème 5.3. La première est plausible mais nécessite quelques réflexions (ici, nous ne donnons pas de démonstration).

Les affirmations du théorème peuvent être observées dans la figure pour l'exemple 2.1 du paragraphe VI.2.

VI.6 Systèmes d'équations différentielles linéaires

Dans ce paragraphe nous considérons des équations différentielles

$$y' = A(x)y + g(x) \tag{6.1}$$

où $A(x)$ est une matrice $n \times n$ et $g(x)$ est un vecteur. On dit que cette équation différentielle est *homogène* si $g(x) \equiv 0$, sinon elle est *inhomogène*.

Théorème 6.1 (existence globale et unicité) *Soit I un intervalle (arbitraire) et supposons que $A(x)$ et $g(x)$ soient des fonctions continues sur I. Alors, le problème de Cauchy $y' = A(x)y + g(x)$, $y(x_0) = y_0$ (avec $x_0 \in I$ et $y_0 \in \mathbb{R}^n$) possède une solution unique sur tout l'intervalle I.*

Démonstration. La fonction $f(x,y) = A(x)y + g(x)$ est continue sur $I \times \mathbb{R}^n$ et satisfait localement une condition de Lipschitz. La solution est donc unique là où elle existe.

Pour démontrer l'existence globale, nous remarquons que les fonctions $y_k(x)$ de la démonstration du théorème 5.3 sont définies sur tout l'intervalle I et pour tout $k \ge 1$. Pour un α arbitraire satisfaisant $x_0 + \alpha \in I$, la démonstration du théorème 5.3 implique l'existence de la solution sur $[x_0, x_0 + \alpha]$. Il suffit de choisir M tel que $\| \int_{x_0}^{x}(A(t)y_0 + g(t))dt\| \le M|x - x_0|$ pour $x_0 \le x \le x_0 + \alpha$. □

Théorème 6.2 (principe de superposition) *Soit I un intervalle et soient $A(x)$, $g_1(x)$, $g_2(x)$ des fonctions continues sur I. Si*

$$\begin{array}{lll} y_1 : I \to \mathbb{R}^n & \textit{est solution de} & y' = A(x)y + g_1(x), \\ y_2 : I \to \mathbb{R}^n & \textit{est solution de} & y' = A(x)y + g_2(x), \end{array}$$

alors $y(x) := c_1 y_1(x) + c_2 y_2(x)$ *est solution de*

$$y' = A(x)y + g(x) \quad \textit{avec} \quad g(x) := c_1 g_1(x) + c_2 g_2(x).$$

Démonstration. Ceci est un exercice très simple. □

VI.6.1 Equations linéaires homogènes

Même si la démonstration du théorème 6.2 est très simple, il a des conséquences très importantes. Nous notons par $y(x, x_0, y_0)$ la solution du problème de Cauchy $y' = f(x, y)$, $y(x_0) = y_0$.

- les solutions de $y' = A(x)y$ forment un espace vectoriel;
- les solutions de $y' = A(x)y$ dépendent linéairement de y_0, c.-à-d.,

  $$y(x, x_0, c_1 y_0 + c_2 z_0) = c_1 y(x, x_0, y_0) + c_2 y(x, x_0, z_0).$$

  La solution de $y' = A(x)y$, $y(x_0) = y_0$ peut donc être écrite sous la forme

  $$y(x, x_0, y_0) = R(x, x_0)y_0, \tag{6.2}$$

  et la matrice $R(x, x_0)$ s'appelle la *résolvante* de l'équation différentielle $y' = A(x)y$. La $i$ème colonne de la matrice $R(x, x_0)$ est solution de $y' = A(x)y$ avec pour valeur initiale $y(x_0) = e_i = (0, \ldots, 0, 1, 0, \ldots 0)^T$;
- si $\Phi(x)$ est une *matrice fondamentale* (aussi appelée *Wronskienne*), c.-à-d., les colonnes de $\Phi(x)$ sont des solutions de $y' = A(x)y$ et $\Phi(x_0)$ est inversible, alors

  $$R(x, x_0) = \Phi(x)\Phi(x_0)^{-1}. \tag{6.3}$$

  En effet, $y(x) = \Phi(x)\Phi(x_0)^{-1}y_0$ est solution de $y' = A(x)y, y(x_0) = y_0$.

Théorème 6.3 (propriétés de la résolvante) *Soit $A(x)$ continue sur un intervalle. Alors, la résolvante de $y' = A(x)y$ satisfait*

i) $R'(x, x_0) = A(x)R(x, x_0)$ *(dérivée par rapport à x)*

ii) $R(x_0, x_0) = I$ *(matrice identité)*

iii) $R(x, x_0) = R(x, x_1)R(x_1, x_0)$

iv) $R(x, x_0)$ *est inversible et* $R(x, x_0)^{-1} = R(x_0, x)$.

Démonstration. Par (6.2) on a que $R'(x, x_0)y_0 = A(x)R(x, x_0)y_0$ et $R(x_0, x_0)y_0 = y_0$ pour tout $y_0 \in \mathbb{R}^n$. Ceci démontre les propriétés (i) et (ii). La propriété (iii) est une conséquence du fait que $R(x, x_0)y_0$ et $R(x, x_1)R(x_1, x_0)y_0$ sont solutions du même problème de Cauchy. Finalement, (iv) suit de (iii) en posant $x = x_0$. □

Exemple 6.4 L'équation différentielle $y'' + y = 0$ peut être écrite sous la forme (en posant $y_1 = y$ et $y_2 = y'$)

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}' = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

et possède comme résolvante

$$R(x, x_0) = \begin{pmatrix} \cos(x - x_0) & \sin(x - x_0) \\ -\sin(x - x_0) & \cos(x - x_0) \end{pmatrix}.$$

Il est intéressant de voir que la propriété (iii) du théorème précédent, c.-à-d.,

$$\begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} \cos\beta & \sin\beta \\ -\sin\beta & \cos\beta \end{pmatrix} = \begin{pmatrix} \cos(\alpha+\beta) & \sin(\alpha+\beta) \\ -\sin(\alpha+\beta) & \cos(\alpha+\beta) \end{pmatrix}$$

avec $\alpha = x - x_1$, $\beta = x_1 - x_0$ et $\alpha + \beta = x - x_0$, est équivalente au théorème d'addition pour $\sin x$ et $\cos x$.

VI.6.2 Equations linéaires inhomogènes

Considérons maintenant l'équation différentielle (6.1) où $g(x)$ n'est pas nulle. Si l'on connait la résolvante de l'équation homogène, on trouve la solution du problème inhomogène par la méthode des variations des constantes.

Théorème 6.5 (variation des constantes) *Soient $A(x)$ et $g(x)$ continues sur un intervalle et soit $R(x, x_0)$ la résolvante de $y' = A(x)y$. Alors, la solution de $y' = A(x)y + g(x)$, $y(x_0) = y_0$ est donnée par*

$$y(x) = R(x, x_0)y_0 + \int_{x_0}^{x} R(x, t)g(t)\, dt.$$

Démonstration. La solution générale de l'équation homogène est $R(x, x_0)c$ avec $c \in \mathbb{R}^n$. L'idée (d'où le nom "variation des constantes") est de chercher une solution de $y' = A(x)y + g(x)$ sous la forme $y(x) = R(x, x_0)c(x)$. Il faut alors que

$$y'(x) = R'(x, x_0)c(x) + R(x, x_0)c'(x) = A(x)R(x, x_0)c(x) + g(x).$$

En utilisant la propriété (i) du théorème 6.3, nous obtenons

$$c'(x) = R(x, x_0)^{-1}g(x) = R(x_0, x)g(x)$$

et par intégration $c(x) = c(x_0) + \int_{x_0}^{x} R(x_0, t)g(t)\, dt$. L'affirmation du théorème découle alors en insérant cette formule pour $c(x)$ dans $y(x) = R(x, x_0)c(x)$. □

Ce résultat montre que, comme dans le cas particulier de dimension 1, la solution générale de $y' = A(x)y + g(x)$ est composée de la solution générale de l'équation homogène et d'une solution particulière de l'équation inhomogène. Il reste alors à discuter le calcul de la résolvante $R(x, x_0)$. Pour le cas $A(x) = A$ (matrice constante), ceci est le sujet du paragraphe suivant. Si $A(x)$ dépend de x, le calcul analytique de la résolvante est rarement possible.

VI.7 Systèmes linéaires à coefficients constants

Considérons le problème de calculer la résolvante de

$$y' = Ay \tag{7.1}$$

où A est une matrice constante d'ordre n (les a_{ij} sont réels ou complexes). Motivés par la résolution de problèmes scalaires, essayons de trouver des solutions de la forme

$$y(x) = e^{\lambda x} v \qquad \text{avec un vecteur } \; v \neq 0\,. \tag{7.2}$$

En insérant (7.2) dans (7.1) nous obtenons $y'(x) = \lambda e^{\lambda x} v = e^{\lambda x} A v$, ce qui est équivalent à $Av = \lambda v$. La fonction de (7.2) est alors une solution de (7.1) si et seulement si λ est une valeur propre de A et v un vecteur propre correspondant.

Cas 1 (A est diagonalisable) Dans cette situation, il existe n vecteurs propres indépendants $v_1, \ldots, v_n$ avec valeurs propres $\lambda_1, \ldots, \lambda_n$. Considérons la matrice

$$\Phi(x) = \Big(e^{\lambda_1 x} v_1, \ldots, e^{\lambda_n x} v_n \Big). \tag{7.3}$$

Cette matrice est inversible pour tout x car $\Phi(0)$ l'est. La résolvante est alors donnée par $R(x, x_0) = \Phi(x)\Phi(x_0)^{-1}$ (voir l'équation (6.3)) ou aussi par $R(x, x_0) = \Phi(x - x_0)\Phi(0)^{-1}$.

Cas 2 (A n'est pas diagonalisable) L'idée est de transformer A sous une forme "plus simple", par exemple sous forme triangulaire (voir le paragraphe II.2) ou sous forme de Jordan (voir le cours "Algèbre I"). On cherche alors une matrice inversible U telle que $U^{-1}AU = S$ possède une telle forme. Avec la transformation $y = Uz$ (et $y' = Uz'$) le système (7.1) devient $z' = Sz$. Pour le cas d'une matrice triangulaire on a

$$\begin{array}{rclllllll}
z_1' & = & s_{11} z_1 & + & s_{12} z_2 & + & \ldots & + & s_{1n} z_n \\
z_2' & = & & & s_{22} z_2 & + & \ldots & + & s_{2n} z_n \\
 & & & & & & \ldots & & \\
z_n' & = & & & & & & & s_{nn} z_n
\end{array}$$

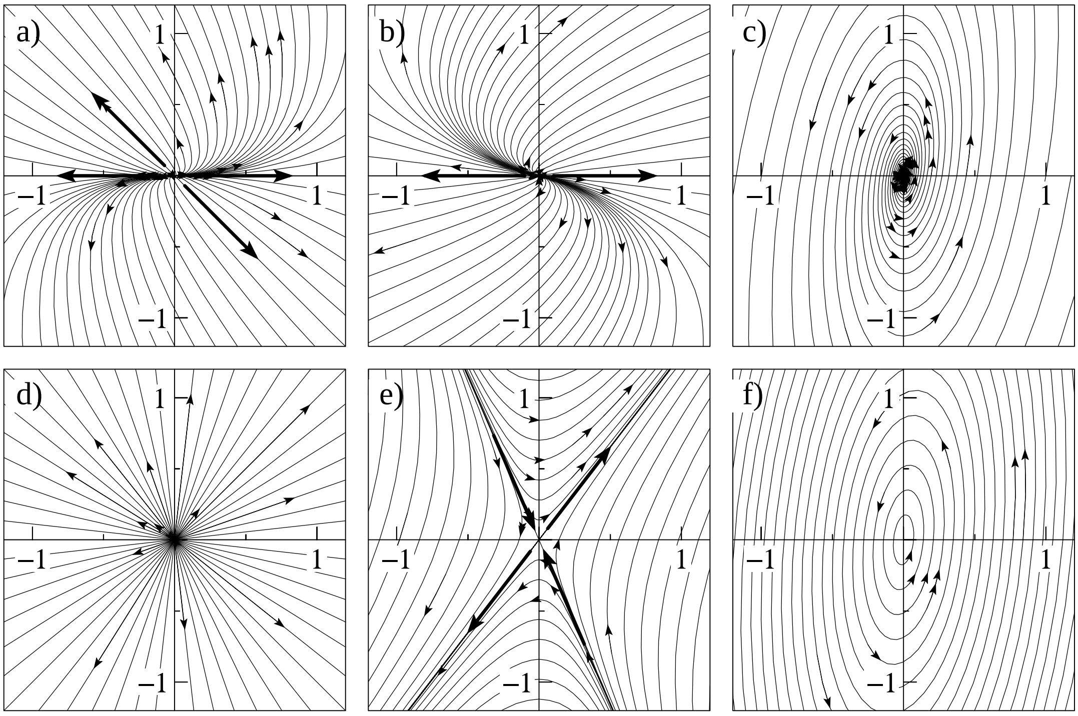
On résoud d'abord l'équation différentielle pour z_n, ensuite celle pour z_{n-1} et à la fin celle pour z_1. La solution de (7.1) est obtenue par $y(x) = Uz(x)$.

Exemple 7.1 Considérons un bloc de Jordan de dimension 4

$$J = \begin{pmatrix} \lambda & 1 & & \\ & \lambda & 1 & \\ & & \lambda & 1 \\ & & & \lambda \end{pmatrix}.$$

Pour la solution de $z' = Jz$ on commence par z_4; on trouve $z_4' = \lambda z_4$ et $z_4(x) = e^{\lambda x} c_4$. L'équation différentielle pour z_3 est $z_3' = \lambda z_3 + e^{\lambda x} c_4$. Sa solution est $z_3(x) = e^{\lambda x} c_3 + x e^{\lambda x} c_4$. De la même façon on calcule $z_2(x)$ et $z_1(x)$. Le résultat est

$$z(x) = e^{\lambda x} \begin{pmatrix} 1 & x & x^2/2! & x^3/3! \\ & 1 & x & x^2/2! \\ & & 1 & x \\ & & & 1 \end{pmatrix} z(0). \tag{7.4}$$



a) $\begin{pmatrix} 1 & -1 \\ 0 & 2 \end{pmatrix}$ b) $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ c) $\begin{pmatrix} 1/3 & -1/3 \\ 2 & 0 \end{pmatrix}$

d) $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ e) $\begin{pmatrix} 1/3 & 1/3 \\ 1 & 0 \end{pmatrix}$ f) $\begin{pmatrix} 1/6 & -1/3 \\ 2 & -1/6 \end{pmatrix}$

Figure VI.7: Solutions de systèmes linéaires en dimension 2

Exemple 7.2 Dans la dimension 2, les solutions $y(x)$ de l'équation différentielle $y' = Ay$ peuvent être dessinées comme courbes avec x comme paramètre. Le comportement des solutions (pour quelques matrices) est illustré dans la Fig. VI.7. Dans les dessins (a), (b) et (e) les directions des vecteurs propres sont indiquées par des flèches.

a) deux valeurs propres distinctes et réelles,
b) une valeur propre de multiplicité deux, seulement un vecteur propre,
c) les valeurs propres sont complexes conjuguées avec $\Re\lambda_i > 0$,
d) une valeur propre de multiplicité deux, mais deux vecteurs propres indépendents,
e) les valeurs propres sont réelles de signe opposé,
f) les deux valeurs propres sont sur l'axe imaginaire.

VI.8 Exercices

Chapter VII

Séries de Fourier

La théorie de ce chapitre nous permet de mieux comprendre toute sorte de phénomènes *périodiques*. Une grande partie de la présentation, des remarques historiques et des figures illustratives a été empruntée (avec permission) de l'excellent polycopié "Analysis II. Pars A" de Gerhardus Wannerus (anno MMI/MMII).

L'étude de la propagation des ondes dans une corde vibrante (Taylor 1713, Joh. Bernoulli 1727, Euler 1739) ou dans l'air (théorie du son de Lagrange 1757) ainsi que l'interpolation de fonctions périodiques en astronomie (Euler 1753) étaient à l'origine de cette théorie. Le livre qui a donné le nom au chapitre est la *Théorie analytique de la Chaleur* de Joseph Fourier. Un premier manuscript a été présenté par Fourier à L'Académie en 1807, un second en 1811 et le livre a finalement été publié en 1822. L'importance de cette œuvre est exprimée par la phrase "FOURIERS *Théorie analytique de la Chaleur* ist die Bibel des mathematischen Physikers" dans un livre de Sommerfeld (1947).

Comme exemple, considérons la digitalisation d'un son (figure VII.1). On a enregistré $22\,000$ impulsions par seconde, dont 1024 sont dessinées (ceci correspond à $1024/22 \approx 46.5$ millisecondes). On est souvent intéressé par l'étude du spectre d'un tel signal, par les fréquences dominantes, par la suppression d'un bruit de fond éventuel, etc.

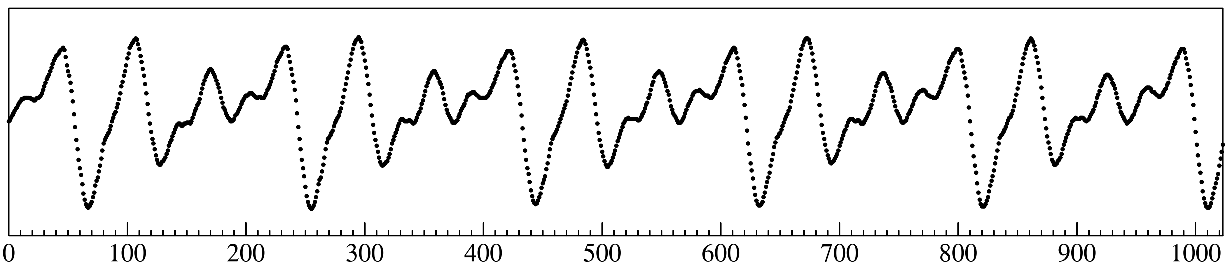


Figure VII.1: Digitalisation du son "o" prononcé par Martin

Aujourd'hui, on utilise des séries de Fourier (sa version discrète FFT, voir le cours "Analyse Numérique", un autre des "Top 10 Algorithms of the 20th Century") et des modifications (wavelets) dans beaucoup d'applications en informatique (compression de sons, compression d'image).

VII.1 Définitions mathématiques et exemples

Fonctions périodiques. Les fonctions $\sin x$, $\cos x$, mais aussi $\sin 2x$, $\cos 5x$ sont des fonctions 2π-*périodiques*, c.-à-d. elles vérifient la relation

$$f(x+2\pi) = f(x) \qquad \text{pour tout } x \in \mathbb{R}.$$

Polynômes trigonométriques. Les combinaisons linéaires de $\sin kx$ et de $\cos kx$ ($k \in \mathbb{Z}$) sont des fonctions 2π-périodiques. Elles sont de la forme

$$\frac{a_0}{2} + \sum_{k=1}^{N} a_k \cos kx + \sum_{k=1}^{N} b_k \sin kx, \tag{1.1}$$

où les a_k et b_k sont des coefficients réels. On appelle (1.1) un *polynôme trigonométrique*.

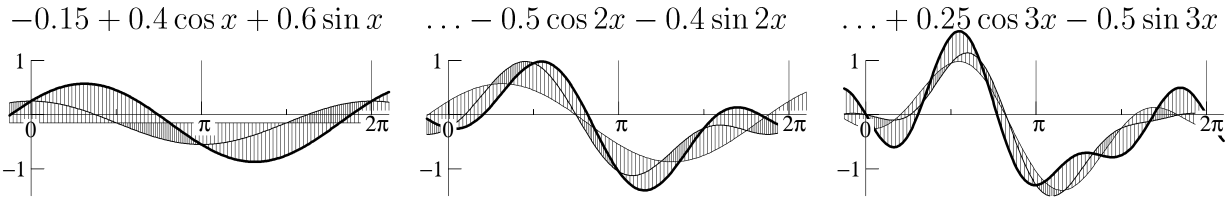


Figure VII.2: Plusieurs polynômes trigonométriques

Formule d'Euler et représentation complexe. La formule d'Euler

$$e^{ix} = \cos x + i \sin x \tag{1.2}$$

nous permet de simplifier l'expression (1.1). En effet, en additionnant et en soustrayant la formule (1.2) et $e^{-ix} = \cos(-x) + i\sin(-x) = \cos x - i \sin x$, on en déduit

$$\cos x = \frac{e^{ix} + e^{-ix}}{2}, \qquad \sin x = \frac{e^{ix} - e^{-ix}}{2i}. \tag{1.3}$$

Le polynôme trigonométrique (1.1) peut donc être écrit sous la forme

$$\sum_{k=-N}^{N} c_k e^{ikx} \tag{1.4}$$

où

$$c_k = \tfrac{1}{2}(a_k - ib_k), \quad c_{-k} = \tfrac{1}{2}(a_k + ib_k) \qquad \text{ou, de manière équivalente,} \qquad a_k = c_k + c_{-k}, \quad b_k = i(c_k - c_{-k}).$$

Avec ces formules, on peut passer de la représentation réelle (1.1) à la représentation complexe (1.4) et vice-versa.

Coefficients de Fourier d'une fonction 2π-périodique. Considérons d'abord un polynôme trigonométrique $f(x) = \sum_{k=-N}^{N} c_k e^{ikx}$, multiplions-le par e^{-ilx} et intégrons de 0 à 2π:

$$\int_0^{2\pi} f(x)\, e^{-ilx}\, dx = \sum_{k=-N}^{N} c_k \int_0^{2\pi} e^{i(k-l)x}\, dx = 2\pi c_l\,,$$

car $\int_0^{2\pi} e^{imx}\, dx = \frac{1}{im} e^{imx}\big|_0^{2\pi} = 0$ pour $m \neq 0$ est égal à 2π pour $m = 0$. On obtient alors

$$\begin{aligned} c_k &= \tfrac{1}{2\pi} \int_0^{2\pi} f(x) e^{-ikx}\, dx \\ a_k &= c_k + c_{-k} = \tfrac{1}{\pi} \int_0^{2\pi} f(x) \cos kx\, dx \\ b_k &= i(c_k - c_{-k}) = \tfrac{1}{\pi} \int_0^{2\pi} f(x) \sin kx\, dx\,. \end{aligned} \tag{1.5}$$

Comme les fonctions sont 2π-périodiques, on peut écrire $\int_{-\pi}^{\pi}$ ou $\int_a^{2\pi+a}$ au lieu de $\int_0^{2\pi}$.

Si $f(x)$ est 2π-périodique, mais pas nécessairement un polynôme trigonométrique, on appelle (1.5) les *coefficients de Fourier* de la fonction $f(x)$.

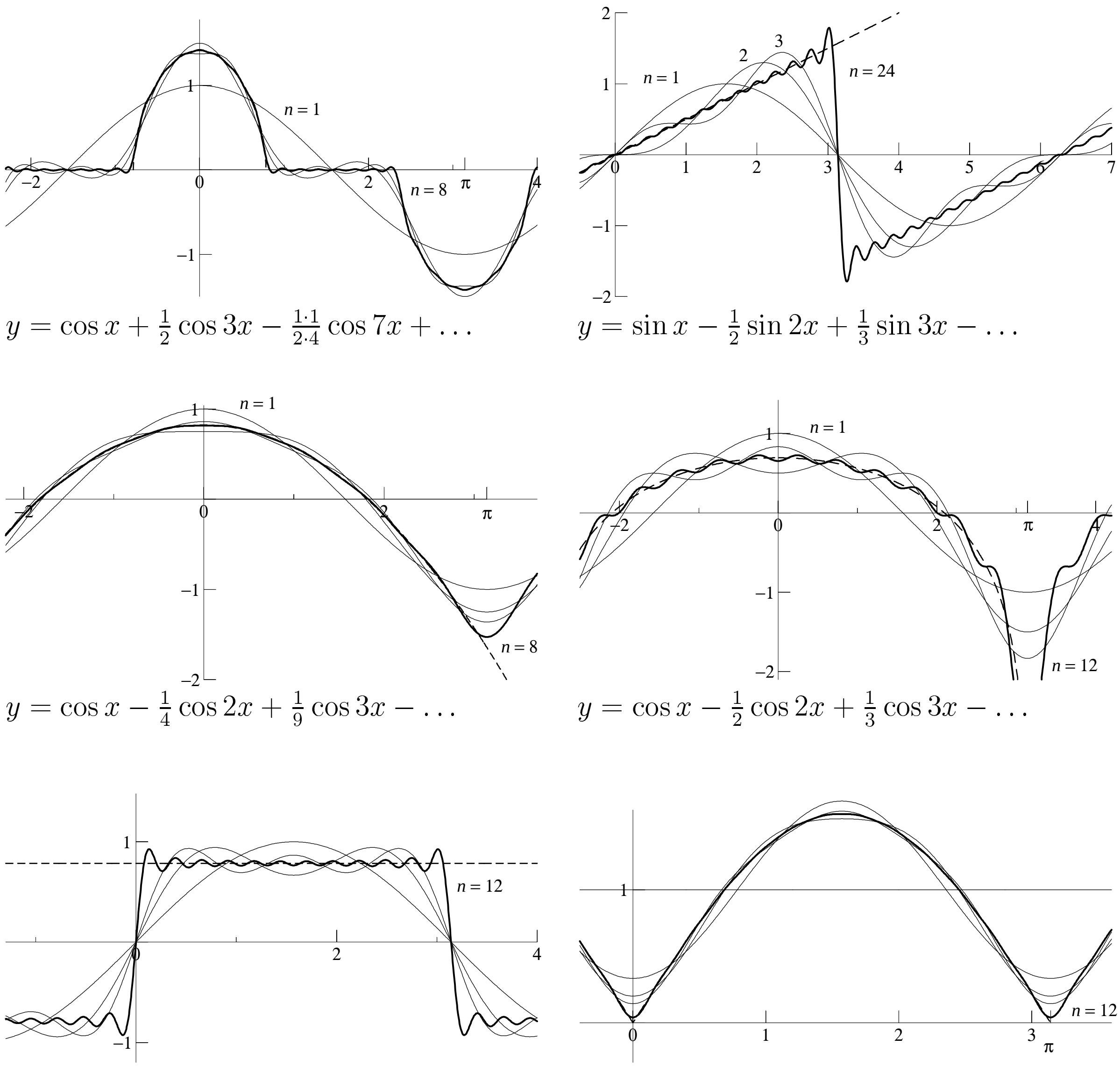
Série de Fourier associée à une fonction périodique. Soit $f(x)$ une fonction 2π-périodique telle que les intégrales dans (1.5) existent. On appelle *“série de Fourier de $f(x)$”* la série

$$\sum_{k\in\mathbb{Z}} c_k e^{ikx} \qquad \text{où} \qquad c_k = \frac{1}{2\pi}\int_0^{2\pi} f(x)e^{-ikx}\,dx \tag{1.6}$$

et on écrit $f(x) \sim \sum_{k\in\mathbb{Z}} c_k e^{ikx}$. La série de Fourier peut également être écrite sous la forme $f(x) \sim a_0/2 + \sum_{k\geq 1} a_k \cos kx + \sum_{k\geq 1} b_k \sin kx$ avec $a_k,\ b_k$ donnés par (1.5).

Pour le moment, on sait seulement qu’on a égalité dans $f(x) = \sum_{k\in\mathbb{Z}} c_k e^{ikx}$ pour des polynômes triogonométriques. On ne sait pas si cette identité reste vraie pour des fonctions 2π-périodiques arbitraires. On ne sait même pas si la série (1.6) converge. Le sujet de ce chapitre est d’étudier cette sorte de questions.

Exemple 1.1 Étudions comment une fonction $f(x)$ est approximée par sa série de Fourier. La figure VII.3 montre six fonctions 2π-périodiques ainsi que plusieurs troncatures de leurs séries de



$y = \cos x + \frac{1}{2}\cos 3x - \frac{1\cdot 1}{2\cdot 4}\cos 7x + \ldots$ $\quad$ $y = \sin x - \frac{1}{2}\sin 2x + \frac{1}{3}\sin 3x - \ldots$

$y = \cos x - \frac{1}{4}\cos 2x + \frac{1}{9}\cos 3x - \ldots$ $\quad$ $y = \cos x - \frac{1}{2}\cos 2x + \frac{1}{3}\cos 3x - \ldots$

$y = \sin x + \frac{1}{3}\sin 3x + \frac{1}{5}\sin 5x + \ldots$ $\quad$ $y = 1 - \frac{2}{1\cdot 3}\cos 2x - \frac{2}{3\cdot 5}\cos 4x - \frac{2}{5\cdot 7}\cos 6x - \ldots$

Figure VII.3: Quelques fonctions 2π-périodiques avec des séries de Fourier tronquées associées

Fourier associées. Les six fonctions sont:

$$f(x) = \begin{cases} \sqrt{2\cos 2x} & \text{si} \quad |x| \le \pi/4 \\ 0 & \text{si} \quad \pi/4 \le |x| \le 3\pi/4 \\ -\sqrt{2\cos 2x} & \text{si} \quad 3\pi/4 \le |x| \le \pi \end{cases} \qquad f(x) = \frac{x}{2} \quad \text{si} \quad |x| < \pi$$

$$f(x) = \frac{\pi^2}{12} - \frac{x^2}{4} \quad \text{si} \quad |x| < \pi \qquad f(x) = \log\Big(2\cos\frac{x}{2}\Big) \quad \text{si} \quad |x| < \pi$$

$$f(x) = \begin{cases} \pi/4 & \text{si} \quad 0 \le x \le \pi \\ -\pi/4 & \text{si} \quad -\pi \le x \le 0 \end{cases} \qquad f(x) = \frac{\pi}{2}|\sin x|$$

Vérifions, par exemple pour la fonction $f(x) = x/2$ ($|x| < \pi$), que la série de Fourier est celle donnée dans le dessin correspondant de la figure VII.3. Un calcul direct donne:

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} \frac{x}{2}\cos kx\,dx = 0 \qquad \text{(parce que } x\cos kx \text{ est impaire),}$$

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} \frac{x}{2}\sin kx\,dx = \frac{1}{2\pi}\Big(-x\,\frac{\cos kx}{k}\Big|_{-\pi}^{\pi} + \frac{1}{k}\int_{-\pi}^{\pi}\cos kx\,dx\Big) = \frac{(-1)^{k+1}}{k}.$$

Remarquons que, en général, on a

- $a_k = 0$ si la fonction $f(x)$ est impaire, c.-à-d. si $f(-x) = -f(x)$ (série de sinus),
- $b_k = 0$ si la fonction $f(x)$ est paire, c.-à-d. si $f(-x) = f(x)$ (série de cosinus).

Exemple 1.2 Étudions encore la fonction du début de ce chapitre qui est la digitalisation d'un son (figure VII.1). Sur l'intervalle comprenant tous les 1024 points elle n'est visiblement pas périodique. Par contre, les premiers 944 points représentent une fonction qui peut être prolongée périodiquement. Sa période est $T = \frac{944}{22\,000}$ secondes. Si on dénote le polygône passant par ces points par $F(t)$ (t en secondes), la fonction $f(x) = F(t)$ avec $x = 2\pi t/T$ devient 2π-périodique et on peut appliquer les formules de ce paragraphe. On cherche donc une représentation

$$F(t) \sim \sum_{k\in\mathbb{Z}} c_k e^{ikx} \qquad \text{avec} \qquad x = \frac{2\pi t}{T}.$$

La figure VII.4 montre les modules de c_k en fonction de k. On les calcule numériquement par FFT, voir le cours "Analyse Numérique". Comme $f(x)$ est réelle, on a $c_{-k} = \overline{c_k}$. On observe que les coefficients de Fourier dominants correspondent tous à des multiples de 5. La fréquence dominante de ce son est donc de $5/T = 5\cdot 22\,000/944 \approx 116.5$ Hz.

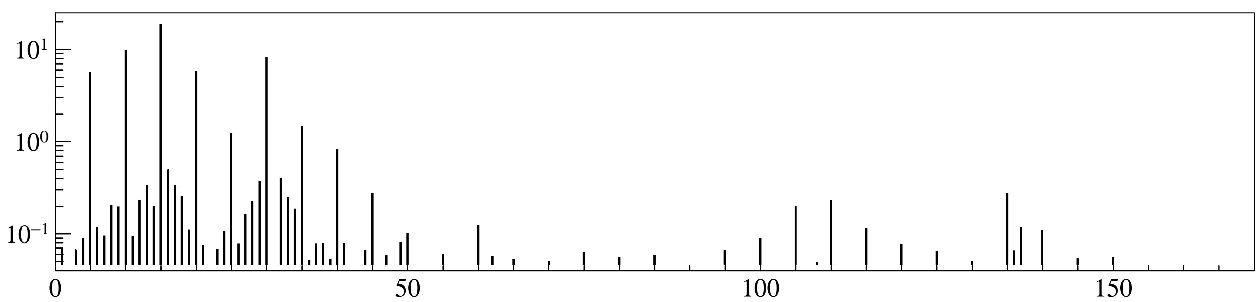


Figure VII.4: Le spectre (valeur absolue de c_k en fonction de k) pour le son de la figure VII.1

k	c_k	$\lvert c_k\rvert$
5	$5.61 - 0.35\,i$	5.62
10	$4.27 - 8.71\,i$	9.70
15	$-13.53 + 12.82\,i$	18.64

k	c_k	$\lvert c_k\rvert$
20	$5.45 - 1.93\,i$	5.78
25	$0.31 - 1.19\,i$	1.23
30	$-7.84 - 2.18\,i$	8.14

Table VII.1: Coefficients de Fourier pour le son de la figure VII.1

Essayons de retrouver le son de la figure VII.1 à partir de son spectre (c.-à-d. à partir des coefficients de Fourier c_k). Quelques valeurs de c_k sont données dans le tableau VII.1. Dans la figure VII.5 nous dessinons quelques séries de Fourier tronquées. D'abord nous prenons uniquement les termes avec $k = +15$ et $k = -15$, c.-à-d. la fonction $c_{-15}e^{-15ix} + c_{15}e^{15ix}$. Ceci donne la fonction pointillée dans la figure VII.5 (un sinus pur). Si on tient en plus compte des coefficients avec $k = \pm 10$ et $k = \pm 30$ on obtient la fonction traitillée, et en ajoutant les termes correspondant à $k = \pm 5$ et $k = \pm 20$ on obtient la courbe solide. Elle est déjà une très bonne approximation du son actuel. On voit qu'avec très peu d'information on peut reconstituer le signal original.

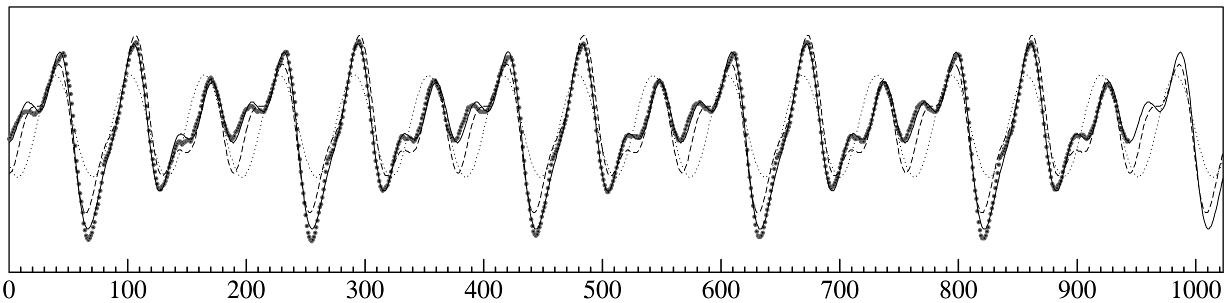


Figure VII.5: Approximation par un polynôme trigonométrique

VII.2 Etude élémentaire de la convergence

Une première approche pour étudier la convergence de la série de Fourier associée à une fonction 2π-périodique consiste à dériver des majorations pour $|c_k|$ et d'utiliser le fait que $|c_k e^{ikx}| \le |c_k|$. Un premier résultat est le suivant:

Lemme 2.1 (lemme de Riemann) *Si $f : [a,b] \to \mathbb{R}$ est intégrable (au sens de Riemann), alors*

$$\lim_{k\to\infty} \int_a^b f(x) \sin kx \, dx = 0 \qquad \textit{et} \qquad \lim_{k\to\infty} \int_a^b f(x) \cos kx \, dx = 0$$

(sans démonstration).

Ce lemme implique que les coefficients de Fourier satisfont toujours

$$a_k \to 0, \quad b_k \to 0 \quad \text{et} \quad c_k \to 0\,.$$

L'hypothèse sur l'intégrabilité de la fonction $f(x)$ est nécessaire car, sinon, les coefficients de Fourier n'existeraient pas. Une definition utile pour estimer les coefficients de Fourier est:

Définition 2.2 Pour une fonction $f : [a,b] \to \mathbb{R}$ on définit sa *variation totale* par

$$V_{[a,b]}f := \sup_{a=x_0<x_1<\ldots<x_n=b} \Big(\sum_{i=0}^{n-1} |f(x_{i+1}) - f(x_i)| \Big).$$

On dit que f est *à variation bornée* si $V_{[a,b]}f < \infty$.

Afin de mieux comprendre cette définition, voici quelques propriétés de fonctions à variation bornée:

- *Si $f : [a,b] \to \mathbb{R}$ est à variation bornée, alors f est intégrable (au sens de Riemann).* Pour une division $D = \{a = x_0 < x_1 < \ldots < x_n = b\}$, la différence entre la grande et la petite somme satisfait (voir les pages 222 et 226 du livre “L’analyse au fil de l’histoire”)

$$\begin{aligned} S(D) - s(D) &= \sum_{i=0}^{n-1} \sup_{u,v\in[x_i,x_{i+1}]} |f(u) - f(v)| \cdot (x_{i+1} - x_i) \\ &\le \sum_{i=0}^{n-1} V_{[x_i,x_{i+1}]} f \cdot (x_{i+1} - x_i) \;\le\; V_{[a,b]} f \cdot \max_{i=0,\ldots,n-1} |x_{i+1} - x_i| \end{aligned}$$

  qui converge vers zéro si $\max_{i=0,\ldots,n-1} |x_{i+1} - x_i| \to 0$.

- *Si f est continûment différentiable, alors f est à variation bornée.* Sur l’intervalle compact $[a,b]$ on a $|f'(x)| \le M$. Le théorème de Lagrange nous donne alors

$$\sum_{i=0}^{n-1} |f(x_{i+1}) - f(x_i)| = \sum_{i=0}^{n-1} |f'(\xi_i)| \cdot |x_{i+1} - x_i| \le M(b-a)\,.$$

- *Une fonction f peut être à variation bornée sans être continue.* Considérons, par exemple, des fonctions en escalier.

- *Une fonction f peut être continue sans être à variation bornée.* Une exemple est la fonction $f(x) = x\,\sin(1/x)$ sur l’intervalle $[0,1]$.

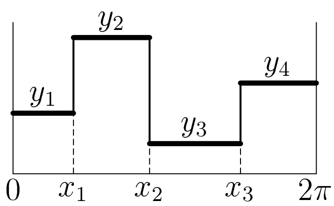
Théorème 2.3 *Si $f : [0,2\pi] \to \mathbb{R}$ est à variation bornée, alors*

$$|c_k| \le \frac{Const}{|k|} \qquad (\text{ de même pour } a_k,\, b_k\,)\,. \tag{2.1}$$

Si $f : \mathbb{R} \to \mathbb{R}$ est 2π-périodique et p fois différentiable avec $f^{(p)}|_{[0,2\pi]}$ à variation bornée, alors

$$|c_k| \le \frac{Const}{|k|^{p+1}}\,. \tag{2.2}$$

Démonstration. Commençons à démontrer l’affirmation (2.1) pour des fonctions en escalier (voir le petit dessin). On peut explicitement calculer les coefficients de Fourier et on obtient



$$\begin{aligned} c_k &= \frac{1}{2\pi}\int_0^{2\pi} s(x)\,e^{-ikx}\,dx \;=\; \frac{1}{2\pi}\sum_{j=1}^{n} y_j \int_{x_{j-1}}^{x_j} e^{-ikx}\,dx \\ &= \frac{1}{2\pi i k}\Big(y_1 + e^{ikx_1}(y_2 - y_1) + e^{ikx_2}(y_3 - y_2) + \ldots - y_n\Big)\,, \\ |c_k| &\le \frac{1}{2\pi k}\Big(V_{[0,2\pi]} s + |s(2\pi) - s(0)|\Big) \;\le\; \frac{Const}{k}\,. \end{aligned}$$

Soit maintenant $f(x)$ une fonction arbitraire à variation bornée. Elle est alors intégrable et, par définition de l’intégrale de Riemann, il existe une fonction en escalier $s(x)$ satisfaisant $V_{[0,2\pi]} s + |s(2\pi) - s(0)| \le V_{[0,2\pi]} f + |f(2\pi) - f(0)|$, tel que $\int_0^{2\pi} s(x)\,e^{-ikx}\,dx$ est arbitrairement proche du coefficient de Fourier $\int_0^{2\pi} f(x)\,e^{-ikx}\,dx$ de $f(x)$. Ceci démontre (2.1).

Si $f(x)$ est une fois différentiable, une intégration par partie donne

$$c_k = \frac{1}{2\pi}\int_0^{2\pi} f(x)e^{-ikx}\,dx = \frac{1}{2\pi}f(x)\frac{e^{-ikx}}{-ik}\Big|_0^{2\pi} + \frac{1}{2\pi ik}\int_0^{2\pi} f'(x)e^{-ikx}\,dx$$

et on peut appliquer (2.1) à $f'(x)$. Plusieurs intégrations par parties donnent (2.2). □

La deuxième et la cinquième fonction de la figure VII.3 sont à variation bornée, mais elles ne sont pas continues sur tout l'intervalle. On comprend alors pourquoi les coefficients de Fourier diminuent comme $Const/k$. La quatrième fonction n'est pas bornée et donc pas à variation bornée. Cela n'empêche pas les coefficients de diminuer aussi comme $Const/k$.

La troisième et la sixième fonction de la figure VII.3 possèdent une première dérivée qui est à variation bornée, d'où un comportement en $Const/k^2$. La première fonction est à variation bornée, mais sa dérivée ne l'est pas. On peut démontrer (par le produit de Wallis) que les coefficients de Fourier se comportent comme $Const/k^{1.5}$.

Concernant la *convergence d'une série de Fourier*, le théorème précédent nous permet de déduire le suivant: si $f(x)$ est 2π-périodique avec une dérivée à variation bornée, alors les coefficients de Fourier satisfont $|c_k| \le Const/k^2$. Le critère de Weierstrass montre donc la convergence uniforme de la série $\sum_k c_k e^{ikx}$. Jusqu'à maintenant on ne sait rien sur la convergence si les coefficients se comportent comme $Const/k$. Dans le cas où la série converge, on ne sait pas encore si elle converge vers la fonction $f(x)$. Ces questions nous occuperont dans le paragraphe suivant.

VII.3 Noyau de Dirichlet et convergence ponctuelle

Pour étudier la convergence de la série de Fourier associée à une fonction $f(x)$, nous considérons les sommes partielles

$$S_n(x) = \sum_{|k|\le n} c_k e^{ikx} \qquad \text{où} \qquad c_k = \frac{1}{2\pi}\int_0^{2\pi} f(x)e^{-ikx}\,dx.$$

Nous nous posons les questions suivantes:

- Pour quelle fonction $f(x)$ et pour quelle valeur de x, la limite $\lim_{n\to\infty} S_n(x)$ existe-t-il?
- Cette limite, est-elle égale à $f(x)$?

Commençons par calculer cette somme partielle:

$$S_n(x) = \sum_{|k|\le n}\frac{1}{2\pi}\int_0^{2\pi} f(t)e^{-ikt}\,dt\cdot e^{ikx} = \int_0^{2\pi} f(t)\Big(\frac{1}{2\pi}\sum_{|k|\le n} e^{ik(x-t)}\Big)dt = \int_0^{2\pi} D_n(x-t)f(t)\,dt$$

où

$$D_n(x-t) := \frac{1}{2\pi}\sum_{|k|\le n} e^{ik(x-t)} \tag{3.1}$$

est le *noyau de Dirichlet*. Des formules plus simples sont données dans le lemme suivant.

Lemme 3.1 *Le noyau de Dirichlet est donné par*

$$D_n(s) = \frac{1}{2\pi}\cdot\frac{\sin{(n+1/2)s}}{\sin s/2} \tag{3.2}$$

et la somme partielle de la série de Fourier satisfait

$$S_n(x) = \int_0^{\pi}\Big(f(x+s)+f(x-s)\Big)D_n(s)\,ds. \tag{3.3}$$

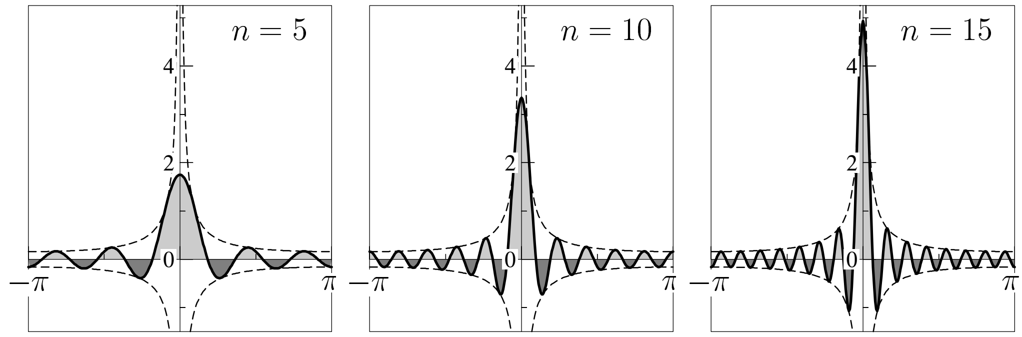


Figure VII.6: Noyaux de Dirichlet $D_n(s)$

Démonstration. En écrivant (3.1) comme une série géométrique, on obtient avec la formule d'Euler

$$\begin{aligned} 2\pi\, D_n(s) &= e^{-ins}\Big(1 + e^{is} + e^{2is} + \ldots + e^{2nis}\Big) = e^{-ins}\,\frac{1-e^{(2n+1)is}}{1-e^{is}} \\ &= \frac{e^{-ins} - e^{i(n+1)s}}{1-e^{is}} = \frac{e^{-i(n+1/2)s} - e^{i(n+1/2)s}}{e^{-is/2} - e^{is/2}} = \frac{\sin\,(n+1/2)s}{\sin s/2}. \end{aligned}$$

La formule pour la somme partielle est donnée par le calcul suivant:

$$\begin{aligned} S_n(x) &= \int_0^{2\pi} f(t) D_n(x-t)\,dt = \int_{x-2\pi}^{x} f(x-s) D_n(s)\,ds = \int_{-\pi}^{\pi} f(x-s) D_n(s)\,ds \\ &= \int_0^{\pi} f(x-s) D_n(s)\,ds + \int_{-\pi}^{0} f(x-s) D_n(s)\,ds = \int_0^{\pi}\Big(f(x+s) + f(x-s)\Big) D_n(s)\,ds. \end{aligned}$$

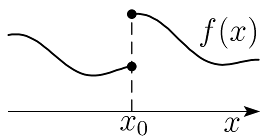
Dans la première ligne on a utilisé la 2π-périodicité de $D_n(s)$ et de $f(x-s)$ et dans la deuxième ligne le fait que $D_n(s) = D_n(-s)$. □

Théorème 3.2 (Dirichlet 1829) *Soit $f(x)$ 2π-périodique, $f|_{[0,2\pi]}$ à variation bornée et supposons qu'en chaque point x_0 de discontinuité les limites $f(x_0-)$ et $f(x_0+)$ existent. Alors la série de Fourier converge pour tout $x_0 \in \mathbb{R}$ et*

- *si f est continue en x_0, on a* $\lim_{n\to\infty} S_n(x_0) = f(x_0)$,
- *si f est discontinue en x_0, on a*

$$\lim_{n\to\infty} S_n(x_0) = \frac{1}{2}\Big(f(x_0-) + f(x_0+)\Big). \tag{3.4}$$

(le premier cas est en fait un cas particulier du deuxième).



Démonstration. Pour simplifier la démonstration, supposons que $f(x)$ soit différentiable dans un voisinage de gauche et de droite de x_0 (voir le petit dessin). Ceci implique que $f(x_0+s) - f(x_0^+) = sg(s)$ pour $s > 0$ avec une fonction $g(s)$ bornée sur $s \geq 0$.

Nous utiliserons la propriété $\int_0^{\pi} D_n(s)\,ds = 1/2$ du noyau de Dirichlet qui est une conséquence de $D_n(-s) = D_n(s)$ et de $\int_{-\pi}^{\pi} D_n(s)\,ds = \int_{-\pi}^{\pi} \frac{1}{2\pi}\sum_{|k|\leq n} e^{iks}\,ds = 1$.

La formule (3.3) nous suggère de considérer l'expression

$$\begin{aligned} \int_0^{\pi} f(x_0+s) D_n(s)\,ds - \frac{f(x_0+)}{2} &= \int_0^{\pi}\Big(f(x_0+s) - f(x_0^+)\Big) D_n(s)\,ds = \int_0^{\pi} s\,g(s) D_n(s)\,ds \\ &= \int_0^{\pi} g(s)\frac{s}{\sin\,(s/2)}\frac{1}{2\pi}\sin\,((n+1/2)s)\,ds \to 0. \end{aligned}$$

La fonction $g(s)s(\sin(s/2))^{-1}$ étant bornée sur l'intervalle $[0,\pi]$, le lemme de Riemann implique que cette expression tend vers zéro pour $n\to\infty$.

De la même manière, on démontre

$$\int_0^\pi f(x_0-s)D_n(s)\,ds-\frac{f(x_0-)}{2}\;\to\;0$$

et en utilisant la formule (3.3) on en déduit (3.4). □

VII.4 Convergence en moyenne quadratique

Rappelons qu'un polynôme trigonométrique de degré n, est une fonction

$$T_n(x)=\sum_{|k|\le n}c_ke^{ikx}\quad\Big(=\frac{a_0}{2}+\sum_{k=1}^n(a_k\cos kx+b_k\sin kx)\Big).$$

Le problème que nous nous posons dans ce paragraphe est le suivant:

Problème. *Soit $f(x)$ une fonction 2π-périodique. Trouver le polynôme trigonométrique de degré n tel que*

$$\int_0^{2\pi}|f(x)-T_n(x)|^2\,dx\;\to\;\min\;.$$

Pour résoudre ce problème, nous le reformulons afin de pouvoir appliquer les techniques du chapitre IV. Nous posons $c_k=\alpha_k+i\beta_k$ et la fonction à minimiser devient

$$F(\alpha_{-n},\beta_{-n},\ldots,\alpha_n,\beta_n):=\int_0^{2\pi}\Big(f(x)-\sum_{|k|\le n}(\alpha_k+i\beta_k)e^{ikx}\Big)\Big(\overline{f(x)}-\sum_{|l|\le n}(\alpha_l-i\beta_l)e^{-ilx}\Big)\,dx.$$

Une condition nécessaire est que toutes les dérivées partielles soient nulles, c.-à-d.

$$\begin{aligned}
\tfrac{\partial F}{\partial\alpha_j}(\ldots) &= -\textstyle\int_0^{2\pi}e^{ijx}\Big(\overline{f(x)}-\sum_{|l|\le n}\overline{c_l}e^{-ilx}\Big)\,dx-\int_0^{2\pi}\Big(f(x)-\sum_{|k|\le n}c_ke^{ikx}\Big)\,e^{-ijx}\,dx\\
&= \ldots=-\Big(\textstyle\int_0^{2\pi}\overline{f(x)}\,e^{ijx}\,dx-2\pi\overline{c_j}\Big)-\Big(\int_0^{2\pi}f(x)e^{-ijx}\,dx-2\pi c_j\Big)=0\\
\tfrac{\partial F}{\partial\beta_j}(\ldots) &= \ldots=-i\Big(\textstyle\int_0^{2\pi}\overline{f(x)}\,e^{ijx}\,dx-2\pi\overline{c_j}\Big)+i\Big(\int_0^{2\pi}f(x)e^{-ijx}\,dx-2\pi c_j\Big)=0\,.
\end{aligned}$$

Comme F est une fonction quadratique, il n'est pas difficile de calculer sa deuxième dérivée (matrice Hessienne) qui est la matrice identité multipliée par 4π. On a alors démontré le théorème suivant:

Théorème 4.1 *La série de Fourier tronquée $S_n(x)$ est parmi tous les polynômes trigonométriques $T_n(x)$ de degré n celui qui minimise l'intégrale*

$$\int_0^{2\pi}|f(x)-T_n(x)|^2\,dx.$$

Théorème 4.2 *Soient $f(x)$ continue par morceaux et $S_n(x)$ la série de Fourier tronquée. Alors*

$$\lim_{n\to\infty}\int_0^{2\pi}|f(x)-S_n(x)|^2\,dx=0\,.$$

Démonstration. Si $f(x)$ est un polynôme trigonométrique, disons de degré 1000, alors $S_n(x) = f(x)$ pour $n \geq 1000$ d'après le théorème précédent et l'affirmation est démontrée.

Pour le cas général, on utilise le théorème de Weierstrass (sans démonstration), qui dit: pour tout $\varepsilon > 0$, il existe un polynôme trigonométrique $T_n(x)$ tel que $|f(x) - T_n(x)| < \varepsilon$ sur $[0, 2\pi]$. Nous ne donnons pas de détails de la démonstration. □

Théorème 4.3 *Soit $f(x)$ intégrable sur $[0, 2\pi]$ et supposons que $\int_0^{2\pi} |f(x)|^2\,dx < \infty$. Les coefficients de Fourier $c_k = \frac{1}{2\pi}\int_0^{2\pi} f(x)e^{-ikx}\,dx$ satisfont alors pour tout n*

$$\sum_{|k|\leq n} |c_k|^2 \leq \frac{1}{2\pi}\int_0^{2\pi} |f(x)|^2\,dx \qquad \textit{(inégalité de Bessel)}\,.$$

Si, en plus, $f(x)$ est continue par morceaux alors

$$\sum_{k\in\mathbb{Z}} |c_k|^2 = \frac{1}{2\pi}\int_0^{2\pi} |f(x)|^2\,dx \qquad \textit{(identité de Parseval)}\,.$$

Démonstration. Calculons

$$\begin{aligned}
0 \;&\leq\; \int_0^{2\pi} |f(x) - S_n(x)|^2\,dx = \int_0^{2\pi} \Big(f(x) - \sum\nolimits_{|k|\leq n} c_k e^{ikx}\Big)\Big(\overline{f(x)} - \sum\nolimits_{|l|\leq n} \overline{c_l} e^{-ilx}\Big)\,dx \\
&= \int_0^{2\pi} |f(x)|^2\,dx - \sum\nolimits_{|k|\leq n} c_k \int_0^{2\pi} \overline{f(x)}\,e^{ikx}\,dx - \sum\nolimits_{|l|\leq n} \overline{c_l} \int_0^{2\pi} f(x)\,e^{-ilx}\,dx + 2\pi \sum\nolimits_{|k|\leq n} |c_k|^2 \\
&= \int_0^{2\pi} |f(x)|^2\,dx - \sum\nolimits_{|k|\leq n} c_k 2\pi \overline{c_k} - \sum\nolimits_{|l|\leq n} \overline{c_l} 2\pi c_l + 2\pi \sum\nolimits_{|k|\leq n} |c_k|^2 \\
&= \int_0^{2\pi} |f(x)|^2\,dx - 2\pi \sum\nolimits_{|k|\leq n} |c_k|^2\,.
\end{aligned}$$

L'identité de Parseval est maintenant une conséquence du théorème précédent. □

VII.5 Exercices